Using Anchor-Based Methods to Determine the Smallest Effect Size of Interest

Farid Anvari

Eindhoven University of Technology; University of Southern Denmark

Daniël Lakens

Eindhoven University of Technology

[accepted for publication at Journal of Experimental Social Psychology]

Farid Anvari is a postdoctoral researcher interested in social psychology, judgment and decision making, meta-science, and measurement, at the Strategic Organization Design group, Department of Marketing and Management, University of Southern Denmark.
Daniel Lakens is an associate professor at the Human-Technology Interaction group interested in applied statistics, research methods, and meta-science.

Correspondence can be addressed to Farid Anvari, Strategic Organization Design group, Department of Marketing and Management, University of Southern Denmark. Email: faridanvari.phd@gmail.com

Acknowledgements

We thank Jarda (Jaroslav) Gottfried for his extensive feedback which helped improve the clarity of our arguments and acknowledging the limitations of the methods. Thanks also to David Funder for literature suggestions.

Disclosures

All data, materials, and code for both studies can be found at https://osf.io/89pcf/. The 2018 cohort in Study 1 was preregistered on the Open Science Framework (https://osf.io/b3z65/) as was Study 2 (https://osf.io/a5pze).

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in both studies.

Protocols for all studies were approved by the Ethical Review Board, Eindhoven University of Technology, approval number 890.

Author Contributions

Both authors jointly generated the idea for the studies and programmed them. F.A. wrote the analysis code and analysed the data, D.L. verified the analyses' accuracy and improved the code for a reproducible report. F.A. wrote the first draft of the manuscript, and both authors critically edited it. Both authors approved the final submitted version.

Conflicts of Interest

No conflicts of interest to declare.

Funding

Farid Anvari was supported by an Australian Endeavour Postdoctoral Fellowship. Daniël Lakens was supported by the Netherlands Organization for Scientific Research VIDI grant 452-17-013 which provided funds for the data collected from Prolific.

Prior Versions

A previous version of the manuscript was posted to the PsyArXiv preprint server (https://psyarxiv.com/syp5a/).

Abstract

Effect sizes are an important outcome of quantitative research, but few guidelines exist that explain how researchers can determine which effect sizes are meaningful. Psychologists often want to study effects that are large enough to make a difference to people's subjective experience. Thus, subjective experience is one way to gauge the meaningfulness of an effect. We propose and illustrate one method for how to quantify the *smallest subjectively experienced difference*—the smallest change in an outcome measure that individuals consider to be meaningful enough in their subjective experience such that they are willing to rate themselves as feeling different—using an anchor-based method with a global rating of change question applied to the positive and negative affect scale. We provide a step-by-step guide for the questions that researchers need to consider in deciding whether and how to use the anchor-based method, and we make explicit the assumptions of the method that future research can examine. For researchers interested in people's subjective experiences, this anchor-based method provides one way to specify a smallest effect size of interest, which allows researchers to interpret observed results in terms of their theoretical and practical significance.

Keywords

Smallest effect size of interest, positive affect, negative affect, practical significance, minimum important difference, subjectively experienced difference, smallest subjectively experienced difference

Using Anchor-Based Methods to Determine the Smallest Effect Size of Interest

Imagine a study that examines whether a simple manipulation will make people feel happier. The study is conducted, and the manipulation causes a statistically significant increase in self-reported positive affect of 0.3 (on a 5-point Likert scale). Is such an increase either theoretically or practically meaningful? To make this evaluation, we need methods to derive empirical benchmarks that can speak to the smallest effect size that has theoretical or practical importance, benchmarks that can serve as the boundary between "interesting" and "uninteresting" effects. One such benchmark is the smallest change that is needed in the outcome measure for people to subjectively notice and report a difference in how they feel—a benchmark based on people's subjective experience.

Psychologists lack clear guidelines for how to interpret the meaningfulness of effect sizes (Fidler, 2002). Some methods exist to determine the smallest effect size that would be theoretically or practically meaningful, but these methods are relatively unknown and not widely used. In this article we first explain why being able to determine a smallest effect size of interest can help to improve the design of experiments and the interpretation of results. Subsequently, we discuss several approaches to determining which effects are deemed meaningful, with a specific focus on one anchor-based method. We provide a step by step guide for how researchers can apply this anchor-based method to determine the smallest effect size of interest, beginning with an outline of the types of research questions this approach is useful for, and noting the considerations that researchers need to keep in mind at each step. Applying the anchor-based method in two studies, we empirically quantify the smallest effect size that individuals subjectively deem to be a meaningful change in positive and negative affect as measured by the Positive and Negative Affect Scale (PANAS; Watson et al., 1988). Finally, in the general

discussion, we make the implicit assumptions of this method explicit, and note how future research can address some of these.

How Determining a Smallest Effect Size of Interest will Improve Research

By examining whether an observed effect size is not just *statistically* significant, but larger than the smallest effect size of interest, researchers can draw conclusions about whether the observed effect is *theoretically* or *practically* significant. This can help to prevent the common misinterpretation of 'statistically significant' as 'meaningful', which is becoming increasingly important given the rise of big data and the uptake of large-scale collaborative projects (e.g., Klein et al., 2014, 2018; Moshontz et al., 2018), where trivially small differences can be statistically significant.

Another important benefit of determining a smallest effect size of interest is that it makes it possible to design an informative and falsifiable study. If a smallest effect size of interest can be determined, researchers can choose a sample size that provides a statistical test with high power to detect it (Albers & Lakens, 2018). Furthermore, it becomes possible to test for, and demonstrate, the absence of an effect that is large enough to be deemed meaningful, by using equivalence tests. For over 60 years, researchers have pointed out the statistical benefits of specifying a range of values that are trivially small (e.g., Hodges & Lehmann, 1954; Nunnaly, 1960), and the benefits of being able to falsify predictions using equivalence tests (Rogers et al., 1993, Lakens et al., 2018). But to be able to reap those benefits, researchers need methods to determine their smallest effect size of interest.

Specifying a smallest effect size of interest is important not only in psychology, but in many other quantitative research fields, including organizational research (Cortina & Landis, 2011; Edwards & Berry, 2010), education research (Hill et al., 2008), communication research (Levine et al., 2008), and clinical and health research (Ferguson, 2009; Kazdin, 1999; King,

2011). In the past, fields have developed either empirical benchmarks for effect sizes that could be considered meaningful, or quantitative methods that can be used to determine minimum thresholds of practical importance for specific research lines.

Methods to Determine a Smallest Effect Size of Interest

Although researchers commonly interpret effect sizes in relation to the benchmarks for small, medium, and large effects suggested by Cohen (1988), these are "arbitrary conventions, recommended for use only when no better basis for estimating the effect size is available" (Cohen, 1988, p. 12). Researchers have attempted to provide more useful benchmarks to be used across fields and studies, based on empirical reviews of effect sizes published in a specific literature (e.g., Ferguson, 2009; Norman et al., 2003; see Funder & Ozer, 2019, for a more detailed and nuanced approach). Although these approaches are useful to interpret the size of an effect relative to that of other effect sizes in the field, they do not quantify which effect sizes are meaningful in specific research lines and, furthermore, there will never be a single answer to the question of which effect size should be considered meaningful. Therefore, it is more appropriate to determine a smallest effect size of interest that is specific to the research question and outcome measure at hand.

One approach in applied research is to perform cost-benefit analyses and determine the size of an effect that could be considered beneficial enough to be worth the costs of an intervention. In intervention studies (e.g., health economics or education research) researchers may determine a smallest effect size of interest by comparing the effect size of the intervention with (1) the change that is expected without implementing the intervention, (2) the size of existing performance gaps (e.g., how much the intervention closes the gaps), and (3) the cost of the intervention, compared to the cost of other interventions (Hill et al., 2008; Torgerson et al., 1995; see also Gruijters & Peters, 2020). However, in more basic research in psychology, costs

and benefits are not easily quantified, as the future applications of the work are less clearly delineated.

But even in basic research, a clear goal can be identified, and it is possible to examine for which effect sizes this goal is met. For example, one question in perception research focusses on the just noticeable difference, or the smallest increase in stimulus intensity that can be reliably noticed by participants. A conceptually related interest in clinical and health research has been the estimation of a minimum threshold of importance for self-reported patient outcomes. The goal is to determine the smallest increase in a relevant outcome measure that is subjectively deemed to be large enough to matter. In the clinical literature the term '*minimal clinically important difference*' is often used (e.g., Chatham et al., 2018) or '*minimally detectable difference'* (Norman et al., 2003), but we will use the umbrella term, minimal important difference (King, 2011).

Several approaches exist that attempt to estimate a minimal important difference. One way to estimate the minimal important difference is to use a clinical *anchor* (Lydick & Epstein, 1993), which functions as a reference to interpret the size of an effect. A common clinical anchor relies on clinician reports, or global ratings, about the extent to which a patient has changed following treatment. For example, clinicians rate whether patients have deteriorated, remained stable, or improved on the domain of interest after treatment. For each group of patients (i.e., those rated as either having deteriorated, remained the same, or improved), the researcher calculates the mean change in the measure of interest that was administered to patients before and after treatment (e.g., Health Related Quality of Life Questionnaire). By referencing the change score (i.e., the difference on the Health Related Quality of Life Questionnaire before and after treatment) to the anchor (i.e., whether the clinician believes the patient has improved or

worsened) researchers can derive an estimate of the minimum important difference—the minimum change in scores on the Health Related Quality of Life Questionnaire that corresponds with what clinicians consider to be a clinically meaningful difference. However, psychologists are more commonly interested in the subjective experiences of people directly, and not in observers' perceptions.

A more patient-centered anchor-based approach to estimate the minimal important difference also uses a global judgment about whether patients have improved or worsened, but asks the patients themselves to provide a subjective global rating of change, which is then used as an anchor (Cuijpers et al., 2014; Devji et al., 2020; Dworkin et al., 2008; Ebrahim et al., 2017; Fleishmann & Vaughan, 2019; Guyatt et al., 2002; Jaeschke et al., 1989; King, 2011; Lydick & Epstein, 1993). In this approach, the construct of interest is also measured at two time-points (T1 and T2), for example before and after an intervention or manipulation. At T2, a global rating of change question asks the extent to which individuals subjectively feel that there has been an increase or decrease since T1 on the construct of interest. For example, at T1 and T2 participants complete the Health Related Quality of Life Questionnaire, and at T2 they answer the global rating of change question concerning the extent to which, since T1, their quality of life has improved or worsened.

Researchers typically use the global rating of change anchor-item to categorize individuals into those who perceive no change, those who perceive a little change (i.e., a little worse or a little better), and those who perceive substantial change (i.e., much worse or much better). The global rating of change item can have a varying number of response options, though some researchers recommend 5 response options to reduce the potential for cutoffs to be subjectively and arbitrarily selected by the researchers (King, 2011; cf. Kamper et al., 2009). The

mean change in scores on the outcome measure of interest (i.e., the Health Related Quality of Life Questionnaire in the example) from T1 to T2 for the individuals who self-report feeling a little worse or a little better is used as the estimate of the minimal important difference. For example, in the studies we report, we focused on the participants who selected "a little less positive" and "a little more positive" for the positive affect dimension of the PANAS, and those who selected "a little less negative" and "a little more negative" for the negative affect dimension. For each of these respective groups, we calculated change in positive affect and negative affect from T1 to T2, to produce estimates of the minimal important difference, which we have termed the smallest subjective experienced difference in positive and negative affect. (For examples of and references to clinical and health research using this approach, see Angst et al., 2001; Button et al., 2015; Cella et al., 2002; Chatham et al., 2018; Devji et al., 2020; Jaeschke et al., 1989; Kounali et al., 2020; Norman et al., 2003; Walters & Brazier, 2005.)

We believe that the global rating of change method can be applied in psychology more generally to determine the smallest effect size of interest, whenever researchers are interested in changes or differences in people's subjective experiences. In what follows, we outline why and how this approach may be useful for basic research in psychology, and then provide a full demonstration of how the method can be used to estimate the smallest effect size of interest for the Positive and Negative Affect Scale (PANAS; Watson et al., 1988).

Estimating the Smallest Subjectively Experienced Difference

Given that many research questions in psychology are concerned with people's subjective experience (how they feel, think, act, and react in various circumstances), it makes sense that the boundary between an "interesting" and an "uninteresting" effect should be also informed by people's subjective experience, rather than by arbitrary thresholds. Psychologists are often interested in effects that are large enough to be subjectively experienced and deemed meaningful

by individuals. For example, many emotion researchers are interested in the subjective experience of emotions (e.g., Campbell-Sills et al., 2006; Coutinho & Cangelosi, 2011; Gross, 1999; Kuppens, 2019; LeDoux, 2014; LeDoux & Hofmann, 2018; Reisenzein, 2009; Troy et al., 2018). By extension, researchers are likely to be interested in effects (e.g., changes or differences in emotion) that are large enough to be subjectively experienced as meaningful. For research questions that center on people's subjective experiences and perceptions, the global rating of change method is one approach that can be used to quantify what people, on average, consider to be a small but subjectively meaningful difference.

Because the global rating of change approach can be used to estimate the smallest difference individuals subjectively perceive as a meaningful change (comparable to, but not to be confused with, the idea of just noticeable differences in psychophysics which was introduced to personality psychology by Ozer, 1993), we refer to this specific estimate of a minimal important difference as the *smallest subjectively experienced difference*. The smallest subjectively experienced difference is the smallest change in an outcome measure that individuals consider to be meaningful enough to rate themselves as feeling different. The more general description of '*smallest effect size of interest*' refers to the smallest effect size that is predicted by theoretical models, considered relevant in daily life, or that is feasible to study empirically (Lakens, 2014). Thus, researchers can use the global rating of change approach to estimate the smallest subjectively experienced difference and, subsequently, use this effect size as a justification for the smallest effect size of interest for relevant research questions.

Implementing the Global Rating of Change Method in Practice

Now we outline, step by step, how researchers can implement the global rating of change anchor method to determine the smallest effect size of interest, including the relevant questions that should be considered at each step (the implementation procedure is visualized in Figure 1).

We demonstrate how to implement the global rating of change approach in practice, with the goal of illustrating its usefulness for basic psychological research. We believe the global rating of change approach has considerable potential, but (in the discussion section) we also reflect on important questions that should be examined in future research if such anchor-based methods become more widely used to determine a smallest effect size of interest.

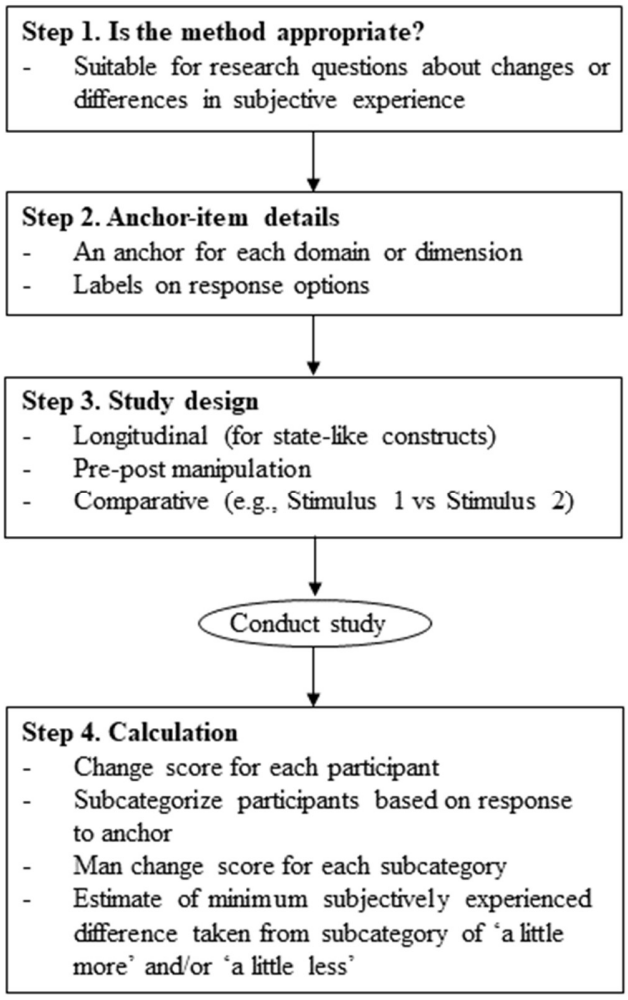

Figure 1. Visual representation of the step by step procedure in deciding on whether and how to use the anchor-based method to estimate the smallest subjectively experienced difference in an outcome measure. Note that the comparative design is described in the discussion section.

The first step requires researchers to consider whether the global rating of change method is appropriate for their line of research. We believe that this approach is useful for research questions that are centered on people's subjective experiences, and when researchers want to draw inferences about whether and how people's feelings, thoughts, and perceptions change over time or are affected by some variable. For example, as we have already noted earlier, the anchor-based approach is suitable for the subjective experience of affect, mood, and emotions.

The second consideration concerns the number of anchor-items to include, as well as the number of response options. For each domain of interest, a unique anchor-item is needed. Thus, for unidimensional outcome measures, a single anchor-item suffices. For outcome measures that are multi-dimensional, an anchor-item is needed for each dimension. For instance, in the studies we report below, the PANAS is a two-dimensional measure, measuring positive affect and negative affect. We therefore included an anchor-item for each, resulting in a total of two anchor-items. For more specific considerations about the anchor-items, such as wording and labelling, we refer readers to Kamper et al. (2009). In the present studies, we opted to use 5 response options, despite some studies suggesting that 7 response options slightly increases reliability (e.g., Kamper et al., 2009), because too many response options (i.e., going beyond 6) can actually make it difficult for participants to differentiate between the options (Simms et al., 2019).

The third step is the study design. For outcomes of interest that vary sufficiently over time, a longitudinal design is acceptable. For example, the subjective experience of emotions is sufficiently state-like, such that how people feel from one day to another will vary. Therefore, researchers can design a longitudinal study (or include an anchor-item into their existing longitudinal design) where the outcome (e.g., affect, mood, or emotion) is measured at two or

more time points and the anchor-item is included at the second and subsequent time points. Similarly, a pre-post design with a manipulation may be used such that the outcome is measured at baseline and then again after a manipulation at which point the anchor-item is also included. The pre-post design is similar to the longitudinal design with the only difference being that there is a manipulation that is expected to create variation in the attribute being measured. (There is also a comparative design that can be used which we detail in the discussion section.)

Studies can be designed for the explicit purpose of determining the smallest subjective experienced difference in the outcome measure, but it might also be possible to include anchor-items in existing study proposals that match the designs outlined above so that the smallest subjectively experienced difference can be estimated. Given that relatively large samples are required to produce precise estimates, a meta-analysis of such studies may provide sufficiently reliable estimates, and such collaborative approaches are recommended.

The fourth step, once the study has been conducted, is to calculate the estimate of the smallest subjectively experienced difference. The calculation is as follows: (i) For each participant, researchers obtain a difference score by subtracting the Time 2 score from the Time 1 score; (ii) participants are subcategorized based on responses on the anchor-item; (iii) for each subcategory, the mean of the difference, obtained from the calculation in (i) above, is calculated; (iv) the mean difference score for participants subcategorized by the anchor-item as either ‘a little less’ or ‘a little more’ on the outcome of interest is taken as the estimate of the smallest subjectively experienced difference. Corresponding standardized effect sizes of interest can also be calculated as can percentage of maximum possible (POMP) score units. Cohen et al. (1999) provide a full description of the benefits of POMP which we summarize very briefly later in this paper. We provide full analytic code to perform such analyses in an online appendix.

At the fourth step, researchers must also consider two other factors in their calculation. If there is symmetry between the mean difference score for the subcategories ‘a little less’ and ‘a little more’, then these may be combined into a single estimate, after first reverse scoring the ‘a little less’ group (i.e., multiplying by -1). If there is asymmetry, such that the difference score for one of these subcategories is considerably larger than the difference score for the other subcategory, then it may be best to report smallest subjectively experienced difference for the ‘a little less’ and ‘a little more’ groups separately. Moreover, researchers may need to consider whether to account for the difference scores of the group of participants (or target object pairs) subcategorized as being ‘the same’, as we have done in the illustration below.

For illustrative purposes, we present two studies in which we used the global rating of change method to determine the smallest subjectively experienced difference and smallest effect size of interest for positive and negative affect, as measured by the PANAS. The PANAS is a widely used measure of positive and negative affect for which researchers have reported difficulty in interpreting how much difference on the scale reflects a meaningful change in affect (e.g., von Leupoldt et al., 2007).

We performed two largely identical studies, both consisting of two independent samples. Because our goal is to illustrate how the global rating of change approach can be used to derive precise estimates of the smallest subjectively experienced difference, we present the results of the combined datasets after detailing the methods for each study. The point estimates and confidence intervals for the ‘little change’ groups across both studies were very consistent, and results for each study and each subsample are presented in the Supplemental Materials. The data and code for both studies can be found on the Open Science Framework (https://osf.io/89pcf/?view_only=ba6c62c915a94873b03295892959d197). We report how we

determined our sample size, all data exclusions, all manipulations, and all measures in both studies.

Study 1

Method

The sampling procedure, methods, and analysis plan for the 2018 cohort of Study 1 were pre-registered on the Open Science Framework (https://osf.io/b3z65/?view_only=dbb79ddde5954f1699cf3ace7724f2bd).

Participants. As part of an assignment related to a psychology lecture on emotions, students completed the PANAS items for course requirements using SurveyMonkey at Time 1 (T1) on Wednesday September 5th (2018 cohort, n = 193) or Wednesday September 4th (2019 cohort, n = 184). At Time 2 (T2), the same students completed the PANAS items on Friday September 7th (2018 cohort; n = 186) or Friday September 6th (2019 cohort; n = 155).[1] The sample size for each cohort was based on the number of students enrolled in the course. We did not include demographic questions for either cohort, but the course from which we drew the 2018 sample had 57% female (43% male) students, mean age 19.4 years (SD = 1.8), and the 2019 course had 54% female (46% male) students, mean age 19.2 years (SD = 2.1). We included only participants with complete responses on all T1 and T2 items. During data cleaning we noticed some students completed the PANAS twice in a row (most likely because of uncertainty about whether answers were submitted correctly), but we only used the first response for each unique student number (N = 316).

Procedure and Measures. At both T1 and T2, participants read that "This scale consists of a number of words that describe different feelings and emotions. Read each item and then

[1] Some participants completed the survey one or two days late.

indicate to what extent you have felt this way today". Participants rated each item (presented in random order) on 5-point Likert scales (from 1 = *very slightly or not at all*, to 5 = *extremely*). Ten items measured positive affect (attentive, interested, alert, excited, enthusiastic, inspired, proud, determined, strong, and active) and ten items measured negative affect (distressed, upset, hostile, irritable, scared, afraid, ashamed, guilty, nervous, jittery). At T2, after responding to the PANAS items, participants also responded to two global rating of change questions, one each for positive and negative emotions. We asked: "Compared to Wednesday, how would you rate the extent of your positive/negative emotions today?". The two global rating of change questions each had 5 response options: much less positive/negative, a little less positive/negative, the same, a little more positive/negative, and much more positive/negative.

Study 2

Method

The sampling procedure, methods, and analysis plan for Study 2 were pre-registered on the Open Science Framework (https://osf.io/a5pze/?view_only=670167b9b3204318b5e2a50c6a63b61f). The study was identical to Study 1, with the exception that to examine the generalizability of our estimates, the time between T1 and T2 was either 2 days (as in Study 2) or 5 days.

Participants. We used the MBESS package in R (Kelley, 2019) to calculate the sample size required for a 95% confidence interval with 0.25 width and an expected estimate of 0.35. This produced a sample size requirement of 500—in the "little change" groups ("little less" and "little more"). After collecting the data from Study 1 approximately 200 participants fell in the "little change" groups based on their response on the anchor question. Because we planned to report the combined results of all samples, we needed a further 300 participants in the "little change" groups after collecting data in Study 2. With this aim, we recruited a total of 550

participants at Time 1, using Prolific (www.prolific.co), requiring participants to be fluent in English and have U.K. nationality. We invited all participants to take part at T2 either 2 days later (n = 275) or 5 days later (n = 275). Excluding participants with incomplete responses and using only the first response for each unique Prolific ID (4 duplicates were removed), we obtained a total of 459 participants at T2 (n_{2days} = 231 and n_{5days} = 228; 74% female, 26% male; age: M = 37.8 years, SD = 12.4, range =18 to 76 years).

Procedure and Measures. The remaining procedures and measures were the same as for Study 1. (The results for the 2- and 5-day delays are presented separately in the Supplemental Materials, with very similar estimates.)

Results of Combined Dataset

The smallest subjectively experienced difference can be presented as a raw score, expressed on the scale that was used to measure it (e.g., 0.3 points on a 5-point scale), as a standardized effect size (e.g., Cohen's d of 0.3), or as the percentage of maximum score units (POMP; Cohen et al., 1999). We recommend reporting estimates in the form of raw differences, standardized effect sizes, and POMP units. Raw scores have the advantage that they do not depend on the standard deviation of the measurements and are perhaps more easily interpreted. Using estimates in the form of raw scores is particularly useful for interpreting results of studies that use the same outcome measure and response options (e.g., a 5-point Likert scale).

Standardized effect sizes have the advantage that they can be compared across instruments and scales. Standardized effect sizes for paired observations can either take the correlation between observations into account (Cohen's d_z) or not (Cohen's d_{av}). Cohen's d_z is used in power analyses and in equivalence tests if the equivalence bounds are set in terms of a standardized effect size. Cohen's d_{av} can in theory be more easily compared across within- and between-subjects designs, although future research should examine whether subjectively

experienced differences in different outcome measures can be assumed to be constant across within and between designs. A limitation of standardized effect sizes is that if the population standard deviation differs across contexts (e.g., the change scores have less variation in a lab experiment compared to a field study), the standardized effect size will differ even though the raw scale difference could be the same. It is important to compare (or even test) estimates of the standard deviation across contexts when examining the generalizability of effects. Depending on the question of interest, either the raw score or POMP estimates might be more relevant across contexts.

Change scores can also be presented as the percent of maximum possible (POMP) units (see Cohen et al., 1999, for a full description). To calculate POMP units, we create sum scores for positive affect and for negative affect for each participant (i.e., the sum of each participant's ratings for the positive affect items and the sum of ratings for the negative affect items). We do this for T1 and T2 separately and then convert these sum scores into POMP units using the formula: ((observed score – min score) / (max score – min score)) * 100; where observed score is the sum score of the participant's ratings; min score is the minimum possible score on the scale (in this case 10 for positive affect and 10 for negative affect; and max score is the maximum possible score on the scale (50 for positive affect and 50 for negative affect in the PANAS). We then calculate the smallest subjectively experienced difference in the same way as for the raw score units. This transformation expresses participants' ratings as a percentage of the maximum possible rating on the scale. POMP units are therefore comparable across measures with a different number of response options (e.g., 5- vs 7-point Likert scales). And, because POMP is independent of the standard deviation, the estimates are more comparable across

contexts where there is a strong expectation that population standard deviation differ substantially.

In sum, raw scores have the advantage of being expressed in the units from the original measurement, whereas standardized effect size estimates can be more readily compared across studies using different measures of the same construct. POMP aides the interpretation of the raw effect size across scales with different response options (Baguley, 2009). Ideally, fields coordinate their use of validated measures, which would reduce the need of standardization. Reporting raw scores, standardized effect sizes, and POMP scores will allow other researchers to use the most appropriate estimates for their purposes.

Figure 2 shows the distribution of change scores, in positive and negative affect, as a function of participants' responses on the global rating of change. Table 1 presents the mean change in positive and negative affect from T1 to T2 (i.e., mean score at T2 minus mean score at T1), standardized effect sizes and their 95% confidence intervals, and the mean change in POMP units, subcategorizing participants based on their responses to the global rating of change items.

To provide evidence for its validity (described in more detail in the discussion section on the assumptions of the method), ratings on the anchor-item should be positively correlated with participants' change scores, given that the anchor-item is intended to measure change in the construct of interest (Devji et al., 2020; Kamper et al., 2009). Indeed, we find some validity evidence: The anchor-item ratings for positive feelings were strongly and positively correlated with PA change scores, $r = .47$ $CI_{95\%}[.41, .52]$; the anchor-item ratings for negative feelings were strongly and positively correlated with NA change scores, $r = .42$ $CI_{95\%}[.36, .48]$. Devji et al. (2020) suggest that a correlation of at least .5 between change scores and the anchor-item is

necessary. There is therefore room for improvement (though of course one should be wary of using arbitrary cut-offs too stringently).

What lends further support for the validity of the anchor-item is that it correlates more strongly with participants' change scores than with their current state T2 scores (Devji et al., 2020), which indeed we find. T2 scores were positively correlated with responses on the anchor-item (PA: r = .37, $CI_{95\%}$[.31, .43]; NA: r = .27, $CI_{95\%}$[.21, .34]). However, the correlations with the change scores were statistically stronger (PA: r_{dif} = .10, $CI_{95\%}$[.04, .16], NA: r_{dif} = .14, $CI_{95\%}$[.07, 0.22]; Zou, 2007). The ideal of anchor-item responses correlating equally but with an opposite sign with T1 and T2 scores is rarely achieved (Devji et al., 2020). In the present data, the anchor-items correlated much more weakly with the T1 scores (PA: r = -.001, $CI_{95\%}$[-.07, .07]; NA: r = -.10, $CI_{95\%}$[-.17, .03]).

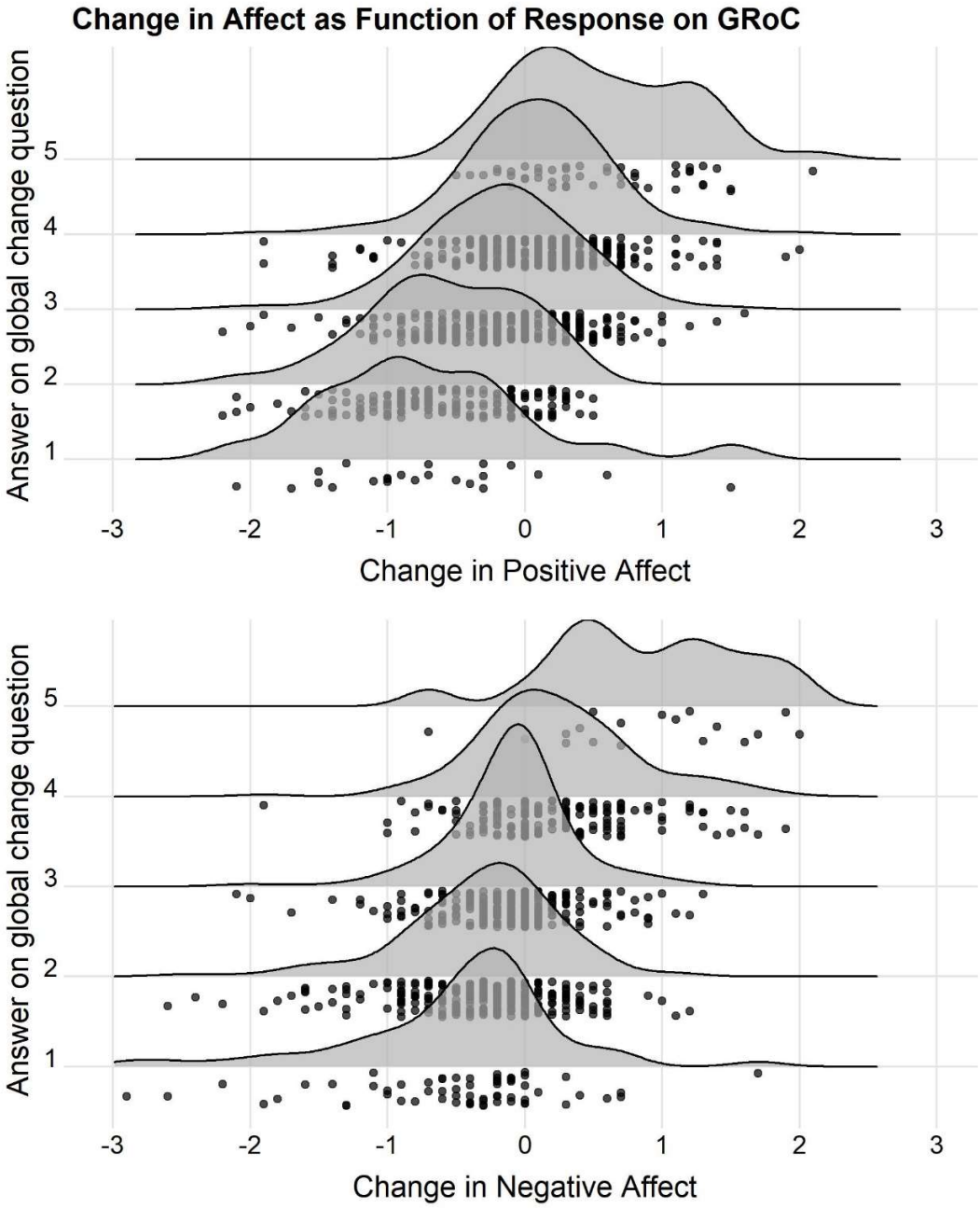

Figure 2. Distribution and individual datapoints of the differences in positive and negative affect as a function of the answer on the global rating of change question. GRoC = Global rating of change. Key for the global rating of change: 1 = much less; 2 = a little less; 3 = the same; 4 = a little more; 5 = much more.

Table 1

Combined Dataset: Means (Standard Deviations) and Mean difference [95% Confidence Intervals] in PANAS scores from T1 to T2, with participants subcategorized based on their responses to the global rating of change question.

	N	T1: M (SD)	T2: M (SD)	Mean Difference	Cohen's d_z	Cohen's d_{av}	POMP
Positive:							
MUCH LESS	23	2.65 (0.88)	1.94 (0.61)	-0.71[-1.05;-0.37]	-0.90[-1.38;-0.41]	-0.94[-1.45;-0.43]	-17.83 [-26.36;-9.29]
A LITTLE LESS	160	3.08 (0.72)	2.51 (0.69)	-0.58[-0.67;-0.49]	-0.99[-1.18;-0.80]	-0.82[-0.98;-0.67]	-14.41[-16.67;-12.14]
THE SAME	264	2.94 (0.81)	2.76 (0.85)	-0.18[-0.25;-0.11]	-0.32[-0.44;-0.19]	-0.22[-0.30;-0.13]	-4.47[-6.18;-2.76]
A LITTLE MORE	281	2.97 (0.75)	3.05 (0.75)	0.08[0.02;0.14]	0.15[0.03;0.26]	0.11[0.02;0.19]	1.99[0.38;3.61]
MUCH MORE	47	3.01 (0.80)	3.56 (0.67)	0.55[0.37;0.72]	0.91[0.57;1.25]	0.74[0.46;1.01]	13.67[9.27;18.07]
Negative:							
MUCH LESS	64	1.88 (0.79)	1.41 (0.49)	-0.48[-0.67;-0.29]	-0.63[-0.89;-0.36]	-0.73[-1.04;-0.42]	-11.95[-16.71;-7.20]
A LITTLE LESS	263	1.95 (0.74)	1.63 (0.60)	-0.32[-0.39;-0.25]	-0.56[-0.69;-0.43]	-0.47[-0.58;-0.36]	-7.94[-9.66;-6.22]
THE SAME	270	1.71 (0.72)	1.61 (0.67)	-0.11[-0.16;-0.05]	-0.24[-0.36;-0.12]	-0.15[-0.23;-0.08]	-2.66[-3.99;-1.32]
A LITTLE MORE	160	1.74 (0.63)	1.97 (0.69)	0.22[0.13;0.31]	0.38[0.22;0.54]	0.34[0.20;0.48]	5.58[3.31;7.84]
MUCH MORE	18	1.91 (0.82)	2.79 (0.89)	0.88 [0.53;1.24]	1.24 [0.61;1.85]	1.03[0.51;1.54]	22.08[13.23;30.94]

Note. Total $N = 775$. T1 = Time 1. T2 = Time 2. Sometimes, the difference in means and standard deviations presented for T1 and T2 do not exactly match with the presented mean difference because they are rounded to the nearest 2 decimals.

First, we see that, somewhat surprisingly, participants who reported feeling ‘the same’ on the global rating of change item actually showed a small decrease in both positive (M = -0.18, SD = 0.57) and negative (M = -0.11, SD = 0.45) PANAS scores. Because the decrease occurs for both positive and negative affect, and we see the expected shifts upward and downward in the ‘a little more/less’ groups for positive and negative affect, relative to the ‘same’ group, this seems to indicate a general decrease with repeated assessment. Such shifts have been observed previously and are referred to as measurement reactivity or the initial elevation bias in subjective reports (Shrout et al., 2018), although the underlying mechanism is not fully understood. Since the smallest subjectively experienced difference can be calculated relative to the ‘no change’ or ‘same’ group (see below) and the initial elevation bias impacts all PANAS scores over the two time points, the relative differences between the ‘little change’ groups and the ‘no change’ group should yield informative estimates.

Some researchers have highlighted the importance of examining whether the little changed groups differ from the unchanged (or same) group, as the absence of a difference casts doubts on the estimated smallest subjectively experienced difference (Hays et al., 2005). As can be inferred from the 95% confidence intervals in Table 1 for both positive and negative affect, the difference scores for people who felt ‘the same’ were statistically different from those who reported feeling a little more or less positive or negative.

Given the shift in evaluations in the ‘no change’ group, possibly due to a general initial elevation bias, it is important calculate the differences between the ‘no change’ and the ‘little more/less’ groups and use this difference as the estimate of how much change is needed, on average, for people to rate themselves as ‘a little more’ or ‘a little less’ positive/negative as opposed to ‘the same’ (Redelmeier et al., 1993, 1997). In our study, for positive affect, those

who said they felt ‘the same’ had change scores that differed from those who said they felt ‘a little more positive’ by an average of 0.26 ($CI_{95\%}$ [0.17, 0.36]) scale points, and from those who said they felt ‘a little less positive’ by -0.39 ($CI_{95\%}$ [-0.28, -0.51]) scale points. For negative affect, those who said they felt ‘the same’ had change scores that differed from those who said that they felt ‘a little more negative’ by an average of 0.33 ($CI_{95\%}$ [0.23, 0.43]) scale points, and from those who felt ‘a little less negative’ by an average of -0.21 ($CI_{95\%}$ [-0.12, -0.30]) scale points.

It is common to combine data from the ‘a little less’ and ‘a little more’ groups into a single estimate of the smallest subjectively experienced difference (after reverse scoring the ‘a little less’ group’s change score by multiplying with -1). This assumes that effects in the positive and negative directions are homogeneous (e.g., have the same size and variance), which is an empirical question for each measurement instrument that the global rating of change method is applied to. We can see from the 95% CI around the change scores reported in the preceding paragraph that the effects are asymmetrical—that is, the CIs of the ‘a little more’ positive/negative groups do not contain the point estimates of the ‘a little less’ positive/negative groups (after reverse scoring the latter), and vice versa. Although the estimates for the ‘a little less’ and ‘a little more’ groups can be combined to produce a single estimate of the smallest subjectively experienced difference for positive (M = 0.31, $CI_{95\%}$ [0.26; 0.36], SD = 0.57) and negative (M = 0.25, $CI_{95\%}$ [0.20; 0.31], SD = 0.57) affect, which can then be used as the smallest effect size of interest for between-groups comparisons in which no baseline control group is available, we believe the observed asymmetry should probably be taken into account when using a smallest subjectively experienced difference in research designs with a baseline control.

The extent to which such asymmetries occur in other psychological measures is a topic for future research. On a related note, one could consider combining the ‘a little more positive’ and ‘a little less negative’ estimates from positive and negative affect, respectively; although these estimates are very similar, we hesitate to do so because, theoretically, positive and negative affect are seen by some as at least partially independent constructs. Therefore, we believe the four estimates presented earlier might be the best level of description for researchers who want to specify a smallest effect size of interest based on the smallest subjectively experienced difference in the PANAS.

The estimate for the smallest subjectively experienced difference is based on the group average. We can therefore expect variability within each group (as visualized in Figure 3), as not every individual will have a change score at or above the group mean, and some people’s change score may contradict their response on the global rating of change item (e.g., an individual who reports feeling a little more positive at T2 can have a negative change score).

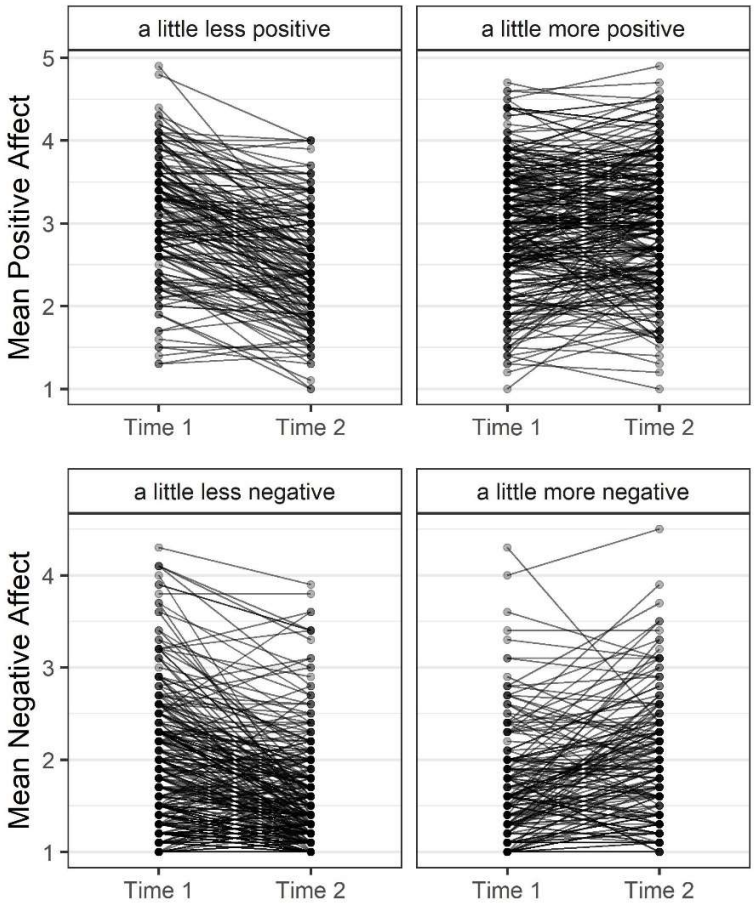

Figure 3. Average self-reported positive and negative affect at T1 and T2, as a function of whether participants reported feeling a little less or more positive and negative at T2, with lines indicating each individual's increase or decrease in scores.

Discussion

The global rating of change approach is an anchor-based method that can be used to determine the smallest subjectively experienced difference for various outcome measures in psychology. We provided a step by step guide that researchers can use, including questions that should be consider each step of the way, to determine the appropriateness of applying the anchor-based method in their research lines and how to do so practically. We presented an

illustrative example by estimating the smallest subjectively experienced difference for positive and negative affect as measured by the PANAS.

We calculated estimates for the relative change, compared to those participants who report feeling 'the same', that led participants to report feeling 'a little more' positive (M = 0.26) and 'a little more' negative (M = 0.33), as well as the relative change that led participants to report feeling 'a little less' positive (M = -0.39) and 'a little less' negative (M = -0.20). These estimates (and their standard deviations) can be used as the smallest effect size of interest in a-priori power analyses, or as the boundaries for an equivalence range when performing an equivalence or minimal effect tests (Lakens, 2014, 2017), for studies that use the PANAS in the population our samples were drawn from (i.e., Dutch university students, and UK participants from Prolific). Alternatively, the smallest subjectively experience difference in POMP units, which can be compared across contexts and measures, were: -14% to feel a little less positive; +2% to feel a little more positive; -8% to feel a little less negative; and +6% to feel a little more negative. We want to stress that researchers should not simply adopt any of these specific estimates as a smallest effect size of interest in all of their future studies. Instead, our recommendation is to use the global rating of change method to determine the smallest effect size of interest for the measure you use in the populations you study.

Since the global rating of change method only requires adding a single anchor-item (for each domain of interest) to longitudinal and pre-post designs, collecting such data should be feasible and easily implemented in existing research designs, such as when researchers examine test-retest reliability—all researchers need to do is include an anchor-item at Time 2 or post manipulation. By combining datasets, fields can establish precise estimates of the smallest subjectively experienced difference, examine their variability, and continue to develop best

practices when implementing anchor-based methods. Although we observed no substantial differences (based on the overlap between point estimates and their confidence intervals) between samples drawn from Dutch university students and U.K. Prolific samples, nor between 2 or 5 days' delay between T1 and T2, it is important not just to improve the precision of estimates, but also to examine their generalizability (King, 2011). Because high-precision estimates require large sample sizes (Maxwell et al., 2008) establishing smallest subjectively experienced differences in specific research fields in psychology would benefit from a coordinated approach to data collection.

The global rating of change method is only one anchor-based approach. Alternative anchor-based approaches exist, such as asking a second individual (e.g., a therapist or close other) to rate the change between T1 and T2 or having people discuss their condition on the domain of interest in pairs before giving global ratings on, for example, how much more or less positive they feel than their partner. This between-person approach has been used in past research to estimate the smallest subjectively perceived difference in walking ability (Redelmeier et al., 1997). Such alternative paradigms might be worthwhile to examine in the future.

The anchor-based approach in this paper can also be adapted to be used for outcome variables that might not vary across time, but which vary across contexts. Here, a comparative design would be more appropriate, with the anchor-item asking people to compare Object 1 with Object 2 on the outcome of interest. For example, to estimate the smallest subjectively experienced difference in facial attractiveness, researchers may have people give attractiveness ratings for a range of face stimuli, and then ask them to make pairwise comparisons between the faces so that participants rate each face relative to the others on attractiveness using the anchor item (i.e., much less attractive, a little less attractive, the same, a little more attractive, much

more attractive). The calculation procedure is similar to that used in longitudinal designs: (i) For each target object pair researchers calculate a difference score by subtracting the ratings for one target object in a pair from the ratings for the other target object; (ii) target object pairs are subcategorized based on responses on the anchor-item; (iii) for each subcategory, the mean of the difference, obtained from the calculation in (i) above, is calculated; (iv) the mean difference score for target object pairs subcategorized by the anchor-item as either 'a little less' or 'a little more' on the outcome of interest is taken as the estimate of the smallest subjectively experienced difference. Hence, the difference in mean ratings between faces that are categorized as a little different from each other (i.e., one face is categorized as either a little less or a little more attractive) provides an estimate of the smallest subjectively experienced difference in facial attractiveness.

To conclude, the anchor-based method can be applied not only to outcome measures that are likely to change over time, but also to outcome measures that are more stable but likely to differ across contexts, in which case the global rating of change would become a global rating of difference.

Assumptions and Future Research Directions

There are several assumptions of the global rating of change anchor-based method that researchers should be aware of (see, Devji et al., 2020; Kamper et al., 2009; King, 2011; Walters & Brazier, 2003). First, the anchor-based approach assumes that researchers are interested in differences that are large enough to be subjectively experienced by people. Sometimes, effect sizes that are too small to be subjectively deemed meaningful by individuals may still be important, and researchers should determine the smallest effect size of interest based on other criteria (such as a cost-benefit analysis, or effect sizes that are theoretically predicted). The

proposed anchor-based approach is primarily suitable for researchers who are interested in effects that people subjectively experience and consider meaningful.

Second, the global rating of change method in longitudinal designs assumes that people can make accurate comparative judgments between what they think or feel now and what they thought or felt when the first instance of the outcome variable was measured (i.e., at Time 1). Some researchers argue that people's retrospective responses on the global rating of change anchor-item may more strongly reflect their present state than their change over time (Norman et al., 1997), whereas responses on the anchor-item should correlate more strongly with people's change scores (Cella et al., 2002; Devji et al., 2020). In our data, people's responses on the anchor-item were more strongly correlated with change scores than with T2 present state scores. If researchers find that participants' responses on the anchor-item are more strongly related to present state than to change scores, then the estimates derived are likely to be suspect and should be used with caution. Nevertheless, given that the global rating of change anchor-item is designed to measure people's *perception* of change in the outcome of interest, a relatively strong correlation between people's responses on the anchor-item and their change scores provides evidence of the anchor-item's validity—that is, the anchor-item is measuring what it purports to measure. As further evidence for the validity of anchor-items, other studies also show significant correlations between responses on global rating of change questions and actual change in the outcome measure of interest (Devji et al., 2020; Kamper et al., 2009).

Third, and related to the assumption pointed out in the preceding paragraph, memory biases may have an impact on estimates derived using the global rating of change method. Thus, it is important to determine whether the anchor-item equally reflects participants' scores at both time points (i.e., T1 and T2) by examining whether the correlation between the anchor-item and

scores at each time point have comparable magnitudes (Norman et al., 1997; Revicki et al., 2008)—global ratings of change which correlate more strongly with T2 scores than with T1 scores indicate that participants are basing their ratings of change more on their present state. In our data, ratings on the anchor-item were far less strongly correlated with T1 scores than with T2 scores. However, the ideal of opposite correlations of equal magnitude between responses on the anchor-item and T1 and T2 scores is rarely achieved in practice, and the anchor-item's validity is nonetheless bolstered by the finding that it correlates more strongly with change scores than with T2 scores (Devji et al., 2020).

Given assumptions 2 and 3, above, researchers should consider that, in longitudinal designs, shorter intervals between time points (as in the present studies) may produce more reliable and valid estimates, and the intervals can be even shorter. Future research can examine the relative validity of estimates across different time intervals. Importantly, in comparative judgment designs, where target object pairs may be compared in real time, validity is likely to be improved and memory biases will not be a problem. When evidence suggests that memory biases exist, as is the case in the present study, we should consider other validity evidence as well. Given the relatively strong correlation between the anchor-item and change scores, as noted in the preceding paragraph, we believe that the present data provide some evidence of validity. If there is evidence of memory bias and little to no evidence of validity (i.e., low or no correlations between the anchor-item and the change scores), then the estimates derived with this approach are extremely suspect.

Although it is important to evaluate and improve the reliability and validity of anchor-based estimates (Devji et al., 2020), and memory biases are known to exist when measuring affect (Napa Scollon et al., 2009), in clinical research, there are typically time delays of *months*

between measurements (e.g., Button et al., 2015) with sufficient accuracy found in delays of around 4 weeks (Devji et al., 2020). We used anchor-items that asked participants to report perceived change over a period of either 2 or 5 days, but even then, it is unreasonable to expect perfect accuracy in recall. Retrospective reports of positive and negative affect are often used in research on subjective well-being, where people are asked to rate how frequently they experienced various feelings in the past 2 to 4 weeks (e.g., Diener et al., 2010; Hudson et al., 2020). Data suggests that such frequency reports are relatively accurate. For example, strong correlations (i.e., *r*s > .60) have been observed between people's daily diary descriptions for how often they felt positive and negative affect each day and their retrospective frequency reports of positive and negative affect over longer periods (i.e., past few days, past week, and past month; Ready et al., 2017; see also Parkinson et al., 1995). Critically, research shows that when people are asked to recall their positive and negative affect at Time 2, *one month* after Time 1, they also show some accuracy (correlations in the range of $r = .90$ for emotions, and $r = .50$ for general mood; Kaplan et al., 2016).

Taken together, the anchor-item is only as valid as its constituent parts—namely, people's ability to (a) recall how they felt at T1, (b) report how they feel at T2, and (c) calculate the difference between these feelings. The evidence noted above suggests that the anchor-item can have sufficient validity, particularly given its strong(er) correlation with change scores (than T2 scores). Nevertheless, the important question about the anchor-item's validity should be examined for each measure to which it will be applied. And although perfectly reliably memory for past psychological states might not be realistic, the anchor-item's reliance on recall is in line with standard practice in the clinical and subjective well-being literatures.

One further caveat pertaining to memory should be mentioned. Researchers who use single-item measures (e.g., a single item measure of positive mood) should consider that people may more readily recall the answer they gave in the past, and use this knowledge (instead of their memory for how they felt) to guide their response on the anchor item. This is likely to be problematic for longitudinal designs such as in the present study, where the delay between T1 and T2 was only a few days. Such memory effects could possibly be counteracted by presenting the anchor item among a set of distractor items.

A fourth assumption is relevant for when researchers want to combine the 'little less' and 'little more' groups to derive a single estimate of the smallest subjectively experienced difference. To combine these subgroups into a single 'little changed' group assumes that the difference in the scores of those who report a little change in the negative and positive direction (i.e., 'little less' and 'little more' groups, respectively) are symmetrical. We believe the assumption that change scores and standard deviations in the two directions are equivalent requires empirical support before scores can be combined. Some researchers calculate the smallest subjectively experienced difference separately for those who improve and those who worsen (e.g., Angst et al., 2001), as we have. Our results suggest that the subgroup of participants who indicated feeling a little more positive (or a little less negative) had a mean change in positive (negative) affect that was considerably lower, in absolute value, than those who said they felt a little less positive (or a little more negative). We believe taking into consideration this asymmetry is the best approach for examining changes in affect as measured by the PANAS.

Fifth, researchers who want to use the estimate of the smallest subjectively experienced difference across different contexts need to consider the assumption that the estimate is

sufficiently invariant across contexts. Although we found similar estimates across 2 and 5 day delays between time points, it is possible that what people may consider ‘a little more positive’ will differ from one context (e.g., Wednesday to Friday) to another (e.g., difference between one image and another). Future research can examine the reliability of the smallest subjectively experienced difference estimates across contexts. For example, researchers can use a pre-post design to measure affect at baseline, introduce a manipulation involving affect-rich stimuli such as images or video clips, then measure affect again and calculate the smallest subjectively experienced difference in this context. If there is large (based on the subjective judgment of the researchers) divergence of the estimates between the pre-post and T1-T2 designs, then researchers will have to determine the smallest subjectively experienced estimate for each context separately.

A sixth assumption is that the smallest subjectively experienced difference will be the same regardless of people’s baseline scores. Although we believe that in the present case (i.e., measuring affect, generally) people’s baseline scores may be of little concern, with the aggregate estimate providing a sufficient level of information, for some research questions the baseline dependency of the estimates should be examined. For example, researchers who want to determine the smallest change in scores that reflect clinically important differences, as with disorders of depression or anxiety, may find baseline scores an important factor (e.g., Button et al., 2015; Kounali et al., 2020).

A seventh assumption is with regards to the reference that participants use to give their global ratings of change. For example, each of the ten items measuring positive affect in the PANAS are assumed to contribute to the construct labelled as positive affect. However, when participants give global ratings of change regarding how positive they feel at T2 relative to T1,

they may not be taking into consideration all of these ten different emotions, perhaps focusing only on one or a few of them, reducing the anchor-item's validity. In our data, for both positive and negative affect, responses on the anchor item were more strongly related to change in the aggregated scores (PA: r = .47; NA: r = .42) than to change in any of the individual items of the scale. For positive affect, change in ratings for the items enthusiastic (r = .34) and excited (r = .33) were the most strongly correlated with the anchor responses, and both of these correlations are statistically smaller than the anchor's correlation with change in the aggregate scores (r_{dif} = .13, $CI_{95\%}$[.08, .19], and r_{dif} = .14, $CI_{95\%}$[.09, 0.20], respectively). Similarly, for negative affect, change in the items upset (r = .36) and irritated (r = .31) had the strongest correlations with the anchor item, both of which were weaker than the anchor's correlation with the aggregated change scores (r_{dif} = .06, $CI_{95\%}$[.004, .11], and r_{dif} = .11, $CI_{95\%}$[.05, 0.17], respectively). Researchers using the anchor item to determine the smallest subjectively experienced difference should similarly examine whether the global judgment of change reflects global change better than change in specific items.

Participants may also be using criteria other than positive affect as part of their judgment of change (Kamper et al., 2009). Indeed, this is reflected in the imperfect correlation between ratings on the anchor-item and participants' change scores, and by the discrepancy between some participants' responses on the anchor-item and their change scores. As discussed above, we believe that the relatively strong correlations between the anchor-items and change scores shows evidence for construct validity. Nevertheless, future research should examine whether and how these correlations might be strengthened.

One fruitful avenue for future validity research would be to develop and examine theoretical models for how people make comparative affect judgments (i.e., compare feelings

from one moment to the next, or compare feelings for one object with another). For example, we can develop models to explain the psychological processes (and other item and person properties) underlying people's responses on the anchor item (e.g., Borsboom et al., 2014; De Boeck & Wilson, 2004; Isager, 2020). Such models would incorporate many of the individual assumptions we've described so far and what's currently known about the psychological processes involved (e.g., for potential biases in retrospective evaluations of affect, see Blome & Augustin, 2015; Thomas & Diener, 1990). The models would then lend themselves to formalization, where all of the assumptions of the measurement process are precisely and transparently specified. Indeed, researchers have already called for the use of formal models in validity research (e.g., Franco et al., 2020). The formal models of measurement can then be used to determine what we would expect to see in an empirical study and to then compare these expectations with actual empirical observations (Franco et al., 2021; Robinaugh et al., 2021). This process of formalization will help improve our understanding of the psychological processes underlying anchor responses and, in so doing, not only help us understand sources of bias and invalidity, but also help us to try and find ways to improve these.

Conclusion

Notwithstanding the above-mentioned assumptions, many of which can be examined in future research, the anchor-based approach we have presented is an improvement over the general lack of consideration of ways to estimate the smallest effect size of interest. To conclude, we believe that the global rating of change method holds promise as one possible approach to estimate the smallest subjectively experienced difference for a variety of psychological measures. Researchers can use these estimates to justify a smallest effect size of interest and interpret the results of their studies in relation to whether observed effects are large enough to be deemed subjectively meaningful by individuals. In the end, we hope anchor-based methods will

help researchers to think more carefully about which effects they consider meaningful in their research

References

American Psychological Association [APA] (2010). *Publication Manual of the American Psychological Association*, 6th Ed. Washington, DC: American Psychological Association.

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187-195. https://doi.org/10.1016/j.jesp.2017.09.004

Angst, F., Aeschlimann, A., & Stucki, G. (2001). Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. *Arthritis & Rheumatism*, *45*(4), 384–391. https://doi.org/10.1002/1529-0131(200108)45:4<384::AID-ART352>3.0.CO;2-0

Baguley, T. (2009). Standardized or simple effect size: What should be reported? British Journal of Psychology, 100(3), 603–617. https://doi.org/10.1348/000712608X377117

Blome, C., & Augustin, M. (2015). Measuring change in quality of life: bias in prospective and retrospective evaluation. *Value in Health*, *18*(1), 110-115. https://doi.org/10.1016/j.jval.2014.10.007

Button, K. S., Kounali, D., Thomas, L., Wiles, N. J., Peters, T. J., Welton, N. J., … Lewis, G. (2015). Minimal clinically important difference on the Beck Depression Inventory - II according to the patient's perspective. *Psychological Medicine*, *45*(15), 3269–3279. https://doi.org/10.1017/S0033291715001270

Campbell-Sills, L., Barlow, D. H., Brown, T. A., & Hofmann, S. G. (2006). Acceptability and suppression of negative emotion in anxiety and mood disorders. *Emotion*, *6*(4), 587-595. https://doi.org/10.1037/1528-3542.6.4.587

Cella, D., Hahn, E. A., & Dineen, K. (2002). Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. *Quality of Life Research*, *11*(3), 207-221. https://doi.org/10.1016/s0149-2918(02)85118-9

Chatham, C. H., Taylor, K. I., Charman, T., Liogier D'ardhuy, X., Eule, E., Fedele, A., … Bolognani, F. (2018). Adaptive behavior in autism: Minimal clinically important differences on the Vineland-II: Adaptive behavior and autism. *Autism Research*, *11*(2), 270–283. https://doi.org/10.1002/aur.1874

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, N.J: L. Erlbaum Associates.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research*, *34*(3), 315-346. https://doi.org/10.1207/S15327906MBR3403_2

Cortina, J. M., & Landis, R. S. (2011). The earth is not round (p=. 00). *Organizational Research Methods*, *14*(2), 332-349. https://doi.org/10.1177/1094428110391542

Coutinho, E., & Cangelosi, A. (2011). Musical emotions: Predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. *Emotion*, *11*(4), 921-937. https://doi.org/10.1037/a0024700

Cuijpers, P., Turner, E. H., Koole, S. L., Van Dijke, A., & Smit, F. (2014). What is the threshold for a clinically relevant effect? The case of major depressive disorders. *Depression and Anxiety*, *31*(5), 374-378. https://doi.org/10.1002/da.22249

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer.

Devji, T., Carrasco-Labra, A., Qasim, A., Phillips, M., Johnston, B. C., Devasenapathy, N., ... & Urquhart, O. (2020). Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. *BMJ*, *369*. https://doi.org/10.1136/bmj.m1714

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D. W., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, *97*(2), 143-156. https://doi.org/10.1007/s11205-009-9493-y

Dworkin, R. H., Turk, D. C., Wyrwich, K. W., Beaton, D., Cleeland, C. S., Farrar, J. T., ... & Brandenburg, N. (2008). Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *The Journal of Pain*, *9*(2), 105-121. https://doi.org/10.1016/j.pain.2009.08.019

Ebrahim, S., Vercammen, K., Sivanand, A., Guyatt, G. H., Carrasco-Labra, A., Fernandes, R. M., ... & Johnston, B. C. (2017). Minimally important differences in patient or proxy-reported outcome studies relevant to children: a systematic review. *Pediatrics*, *139*(3), e20160833, 1-16. https://doi.org/10.1542/peds.2016-0833

Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods*, *13*(4), 668-689. https://doi.org/10.1177/1094428110380467

Ferguson, C. J. (2009). An effect size primer: a guide for clinicians and researchers. *Professional Psychology: Research and Practice 40*(5), 532-538. https://doi.org/10.1037/a0015808

Fidler, F. (2002). The fifth edition of the APA Publication Manual: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement*, *62*(5), 749-770. https//doi.org/10.1177/001316402236876

Fleischmann, M., & Vaughan, B. (2019). Commentary: statistical significance and clinical significance-a call to consider patient reported outcome measures, effect size, confidence interval and minimal clinically important difference (MCID). *Journal of Bodywork and Movement* Therapies. 23(4), 690-694. https://doi.org/10.1016/j.jbmt.2019.02.009

Franco, V. R., Laros, J. A., & Wiberg, M. (2021). *How to Think Straight About Psychometrics: Improving Measurement by Identifying its Assumptions*. PsyArXiv. https://doi.org/10.31234/osf.io/5sjeu

Franco, V. R., Wiberg, M., & Laros, J. A. (2020). Situational optimization function analysis: An ideal performance analysis inspired on Lewin's Equation. PsyArXiv. https://doi.org/10.31234/osf.io/t7uex

Funder, D. C., & Ozer, D. J. (in press). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156-168. https://doi.org/10.1177/2515245919847202

Gross, J. J. (1999). Emotion regulation: Past, present, future. *Cognition & Emotion*, *13*(5), 551-573. https://doi.org/10.1080/026999399379186

Gruijters, S. L., & Peters, G. J. Y. (2020). Meaningful change definitions: Sample size planning for experimental intervention research. *Psychology & Health*, 1-16. https://doi.org/10.1080/08870446.2020.1841762

Guyatt, G. H., Osoba, D., Wu, A. W., Wyrwich, K. W., Norman, G. R., & Clinical Significance Consensus Meeting Group. (2002, April). Methods to explain the clinical significance of

health status measures. In *Mayo Clinic Proceedings, 77*(4), 371-383). https://doi.org/10.4065/77.4.371

Hays, R. D., Farivar, S. S., & Liu, H. (2005). Approaches and Recommendations for Estimating Minimally Important Differences for Health-Related Quality of Life Measures. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, *2*(1), 63–67. https://doi.org/10.1081/COPD-200050663

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172-177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Hodges Jr, J. L., & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society: Series B (Methodological)*, *16*(2), 261-268. https://doi.org/10.1111/j.2517-6161.1954.tb00169.x

Hudson, N. W., Anusic, I., Lucas, R. E., & Donnellan, M. B. (2020). Comparing the reliability and validity of global self-report measures of subjective well-being with experiential day reconstruction measures. *Assessment*, *27*(1), 102-116. https://doi.org/10.1177/1073191117744660

Isager, P. M. (2020). *Test validity defined as d-connection between target and measured attribute: Expanding the causal definition of Borsboom et al. (2004)*. PsyArXiv https://doi.org/10.31234/osf.io/btgsr

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*, *10*(4), 407-415. https://doi.org/10.1016/0197-2456(89)90005-6

Kamper, S. J., Maher, C. G., & Mackay, G. (2009). Global rating of change scales: A review of strengths and weaknesses and considerations for design. *Journal of Manual & Manipulative Therapy*, *17*(3), 163-170. https://doi.org/10.1179/jmt.2009.17.3.163

Kaplan, R. L., Levine, L. J., Lench, H. C., & Safer, M. A. (2016). Forgetting feelings: Opposite biases in reports of the intensity of past emotion and mood. *Emotion*, *16*(3), 309. http://dx.doi.org/10.1037/emo0000127

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology 67*(3), 332-339. https://doi.org/10.1037/0022-006x.67.3.332

Kelley, K. (2019). MBESS: The MBESS R Package. R package version 4.5.1. https://CRAN.R-project.org/package=MBESS

King, M. T. (2011). A point of minimal important difference (MID): a critique of terminology and methods. *Expert Review of Pharmacoeconomics & Outcomes Research*, *11*(2), 171–184. https://doi.org/10.1586/erp.11.9

Kounali, D., Button, K. S., Lewis, G., Gilbody, S., Kessler, D., Araya, R., ... & Lewis, G. (2020). How much change is enough? Evidence from a longitudinal study on depression in UK primary care. *Psychological Medicine*, 1-8. https://doi.org/10.1017/S0033291720003700

Kuppens, P. (2019). Improving theory, measurement, and reality to advance the future of emotion research. *Cognition and Emotion*, *33*(1), 20-23. https://doi.org/10.1080/02699931.2018.1536037

Kvam, A. K., Fayers, P., & Wisloff, F. (2010). What changes in health-related quality of life matter to multiple myeloma patients? A prospective study. *European Journal of Haematology, 84*(4), 345–353. https://doi.org/10.1111/j.1600-0609.2009.01404.x

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355-362. https://doi.org/10.1177/1948550617697177

Lakens, D. (2019, April 9). The practical alternative to the p-value is the correctly used p-value. https://doi.org/10.31234/osf.io/shm8v

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259-269. https://doi.org/10.31234/osf.io/v3zkt

LeDoux, J. E. (2014). Coming to terms with fear. *Proceedings of the National Academy of Sciences*, *111*(8), 2871-2878. https://doi.org/10.1073/pnas.1400335111

LeDoux, J. E., & Hofmann, S. G. (2018). The subjective experience of emotion: a fearful view. *Current Opinion in Behavioral Sciences*, *19*, 67-72. https://doi.org/10.1016/j.cobeha.2017.09.011

Levine, T. R., Weber, R., Park, H. S., & Hullett, C. R. (2008). A communication researchers' guide to null hypothesis significance testing and alternatives. *Human Communication Research*, *34*(2), 188-209. https://doi.org/10.1111/j.1468-2958.2008.00318.x

Lydick, E., & Epstein, R. S. (1993). Interpretation of quality of life changes. *Quality of life Research*, *2*(3), 221-226. https://doi.org/10.1007/bf00435226

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology, 59*(1), 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of Changes in Health-related Quality of Life: The Remarkable Universality of Half a Standard Deviation. *Medical Care*, *41*(5), 582–592. https://doi.org/10.1097/01.MLR.0000062554.74615.4C

Norman, G. R., Stratford, P., & Regehr, G. (1997). Methodological problems in the retrospective computation of responsiveness to change: The lesson of Cronbach. *Journal of Clinical Epidemiology*, *50*(8), 869-879. https://doi.org/10.1016/s0895-4356(97)00097-8

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, *20*(4), 641-650. https://doi.org/10.1177/001316446002000401

Parkinson, B., Briner, R. B., Reynolds, S., & Totterdell, P. (1995). Time frames for mood: Relations between momentary and generalized ratings of affect. *Personality and Social Psychology Bulletin*, *21*(4), 331-339. https://doi.org/10.1177/0146167295214003

Ozer, D. J. (1993). Classical psychophysics and the assessment of agreement and accuracy in judgments of personality. *Journal of Personality*, *61*(4), 739-767. https://doi.org/10.1111/j.1467-6494.1993.tb00789.x

Ready, R. E., Weinberger, M. I., & Jones, K. M. (2007). How happy have you felt lately? Two diary studies of emotion recall in older and younger adults. *Cognition and Emotion*, *21*(4), 728-757. https://doi.org/10.1080/02699930600948269

Redelmeier, D. A. (1993). Assessing the clinical importance of symptomatic improvements. An illustration in rheumatology. *Archives of Internal Medicine*, *153*(11), 1337–1342. https://doi.org/10.1001/archinte.153.11.1337

Redelmeier, D A, Bayoumi, A. M., Goldstein, R. S., & Guyatt, G. H. (1997). Interpreting small differences in functional status: The Six Minute Walk test in chronic lung disease patients. *American Journal of Respiratory and Critical Care Medicine, 155*(4), 1278-1282. https://doi.org/10.1164/ajrccm.155.4.9105067

Redelmeier, Donald A., Guyatt, G. H., & Goldstein, R. S. (1996). On the debate over methods for estimating the clinically important difference. *Journal of Clinical Epidemiology*, *49*(11), 1223–1224. https://doi.org/10.1016/S0895-4356(96)00208-9

Reisenzein, R. (2009). Emotional experience in the computational belief–desire theory of emotion. *Emotion Review*, *1*(3), 214-222. https://doi.org/10.1177/1754073909103589

Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology*, *61*(2), 102-109. https://doi.org/10.1016/j.jclinepi.2007.03.012

Robinaugh, D., Haslbeck, J., Ryan, O., Fried, E. I., & Waldorp, L. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691620974697

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553-565. https://doi.org/10.1037/0033-2909.113.3.553

Scollon C. N., Prieto C. K., Diener E. (2009) Experience sampling: Promises and pitfalls, strength and weaknesses. In: Diener E. (eds) *Assessing Well-Being*. Social Indicators Research Series, vol 39. Springer, Dordrecht. https://doi.org/10.1007/978-90-481-2354-4_8

Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavél, F. D., Iida, M., Gleason, M. E. J., Xu, J. H., & Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences*, *115*(1), E15–E23. https://doi.org/10.1073/pnas.1712277115

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, *31*(4), 557. http://dx.doi.org/10.1037/pas0000648

Thomas, D. L., & Diener, E. (1990). Memory accuracy in the recall of emotions. *Journal of Personality and Social Psychology*, 59(2), 291–297. http://dx.doi.org.dianus.libr.tue.nl/10.1037/0022-3514.59.2.291

Torgerson, D. J., Ryan, M., & Ratcliffe, J. (1995). Economics in sample size determination for clinical trials. *QJM: An International Journal of Medicine*, *88*(7), 517-521. https://doi.org/10.1093/oxfordjournals.qjmed.a069095

Troy, A. S., Shallcross, A. J., Brunner, A., Friedman, R., & Jones, M. C. (2018). Cognitive reappraisal and acceptance: Effects on emotion, physiology, and perceived cognitive costs. *Emotion*, *18*(1), 58-74. https://doi.org/10.1037/emo0000371

von Leupoldt, A., Taube, K., Schubert-Heukeshoven, S., Magnussen, H., & Dahme, B. (2007). Distractive auditory stimuli reduce the unpleasantness of dyspnea during exercise in patients with COPD. *Chest*, *132*(5), 1506–1512. https://doi.org/10.1378/chest.07-1245

Walters, S. J., & Brazier, J. E. (2003). What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. *Health and Quality of Life Outcomes*, *1*(1), 1-8. https://doi.org/10.1186/1477-7525-1-4

Watson, D., Anna, L., & 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. https://doi.org/10.1037/0022-3514.54.6.1063

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*(4), 399–413. https://doi.org/10/fmb3nm