Running head: SMALLEST EFFECT SIZE OF INTEREST

Abstract

Psychology researchers have few guides that explain how they can quantify which effect sizes are theoretically or practically meaningful, yet they often want to study effects that are large enough to matter. For example, when studying interventions aimed at improving subjectively experienced positive affect, the effect size should be large enough for individuals to feel more positive. We illustrate one approach to quantify the *minimum 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. We use an anchor-based method with a global rating of change question, which originates from clinical research, to estimate the minimum subjectively experience difference for positive and negative affect as measured by the positive and negative affect scale. For researchers interested in people’s subjective experiences and conscious perceptions, estimating the minimum subjectively experienced difference provides one way to specify a smallest effect size of interest (SESOI). Being able to specify the smallest effect size that is considered meaningful improves null hypothesis significance testing by allowing researchers to interpret studies’ results in terms of their theoretical and practical significance.

Keywords

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

# Using an Anchor-Based Method 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 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 for self-reported measures of subjective experience. One benchmark we might be interested in is the minimum change that is needed in the outcome measure for people to subjectively notice and report a difference in how they feel.

Psychologists lack clear guidelines for how to interpret the size of effects (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. 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).

# How Determining a SESOI Improves Research

By examining whether an observed effect size is not just *statistically* significant, but larger than the smallest effect size of interest (SESOI), 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, Ratliff et al., 2014; Klein, Vianello et al., 2018; Moshontz et al., 2018), where trivially small differences can be statistically significant.

Another important benefit of determining a SESOI is that it makes it possible to design an informative and falsifiable study. If a SESOI can be determined, it is possible to choose a sample size that provides high power for the smallest effect size that is deemed meaningful (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, Howard, & Vessey, 1993, Lakens, Scheel, & Isager, 2018). But to be able to reap those benefits, researchers need methods to determine their SESOI.

**Methods to Determine a Smallest Effect Size of Interest**

Specifying a SESOI 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, Bloom, Black, & Lipsey, 2008), communication research (Levine et al., 2008), and clinical psychology (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.

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 which effect size should be considered meaningful..

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 SESOI by comparing the effect size of the intervention with (1) the change that is expected when implementing the intervention, (2) the cost 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). 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 more basic research, a clear goal can be identified, and it is possible to examine at 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, Sloan, & Wyrwich, 2003), but we will use the umbrella term, minimal important difference (King, 2011).

Several approaches exist that attempt to estimate a minimal important difference (MID). One approach for estimating the MID 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; HRQoLQ). By referencing the change score (i.e., the difference on HRQoLQ before and after treatment) to the anchor (i.e., whether the clinician believes patients have improved or worsened) researchers can derive an estimate of the MID—the minimum change in scores on the HRQoLQ that a clinician considers to be a clinical meaningful difference. However, psychologists are more commonly interested in the subjective experiences of people directly, and not in the perceived experiences by observers.

A more patient-centered anchor-based approach to estimate the MID also uses a global rating 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; 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 (HRQoLQ), 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 (GRoC) 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 GRoC item usually has 5 ordinal response options, which is recommended because it reduces the potential for cutoffs to be subjectively and arbitrarily selected by the researchers (King, 2011). The mean change in scores on the outcome measure of interest (i.e., the HRQoLQ 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 MID. (For clinical/health research using this approach, see Angst et al., 2001; Button et al., 2015; Jaeschke et al., 1989; Walters & Brazier, 2005.)

We believe that the GroC 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 SESOI for the Positive and Negative Affect Scale (PANAS; Watson et al., 1988).

**Estimating the Minimum Subjectively Experienced Difference and SESOI**

Psychologists are often interested in effects that are large enough to be subjectively experienced and deemed meaningful b individuals. For example, many emotion researchers are interested in the subjective experience of emotions (e.g., Campbell-Sills, Barlow, Brown, & Hofmann, 2006; Coutinho & Cangelosi, 2011; Gross, 1999; Kuppens, 2019; LeDoux, 2014; LeDoux & Hofmann, 2018; Reisenzein, 2009; Troy, Shallcross, Brunner, Friedman, & Jones, 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 GRoC 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 GRoC 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 difference in psychophysics which was introduced to personality psychology by Ozer, 1993), we refer to this specific estimate of a minimal important difference as the *minimum* s*ubjectively experienced difference*. The minimum 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 GRoC approach to estimate the minimum subjectively experienced difference and, subsequently, use this effect size as a justification for the SESOI for relevant research questions.

**Implementing the GRoC Method in Practice**

We demonstrate how to implement the GRoC approach in practice, with the goal to illustrate its usefulness in psychological research. We believe the GRoC approach has considerable potential, but we also reflect on important questions that should be examined in future research if anchor based methods become more widely used to determine a smallest effect size of interest. We use the GRoC method to determine the minimum subjectively experienced difference and SESOI 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, Taube, Schubert-Heukeshoven, Magnussen, & Dahme, 2007).

We performed two largely identical studies, both consisting of two independent samples. Because our goal is to illustrate how the GRoC approach can be used to derive precise estimates of the minimum 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.

**Study 1**

**Method**

**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]](#footnote-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 with a mean age of 19.4 years (*SD* = 1.8) ranging from 17 to 26 years (demographics for the 2019 cohort were not available). We included only participants with complete responses on all T1 and T2 items. During data cleaning we saw some students completed the PANAS twice in a row (most likely because of uncertainty 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 indicate to what extent you have felt this way today”. Participants rated each item (presented in random order) on 5-point Likert-type 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, participants also responded to two GRoC questions, one for positive emotions and one for negative emotions. We asked: “Compared to Wednesday, how would you rate the extent of your positive [negative] emotions today?”. The two GRoC 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**

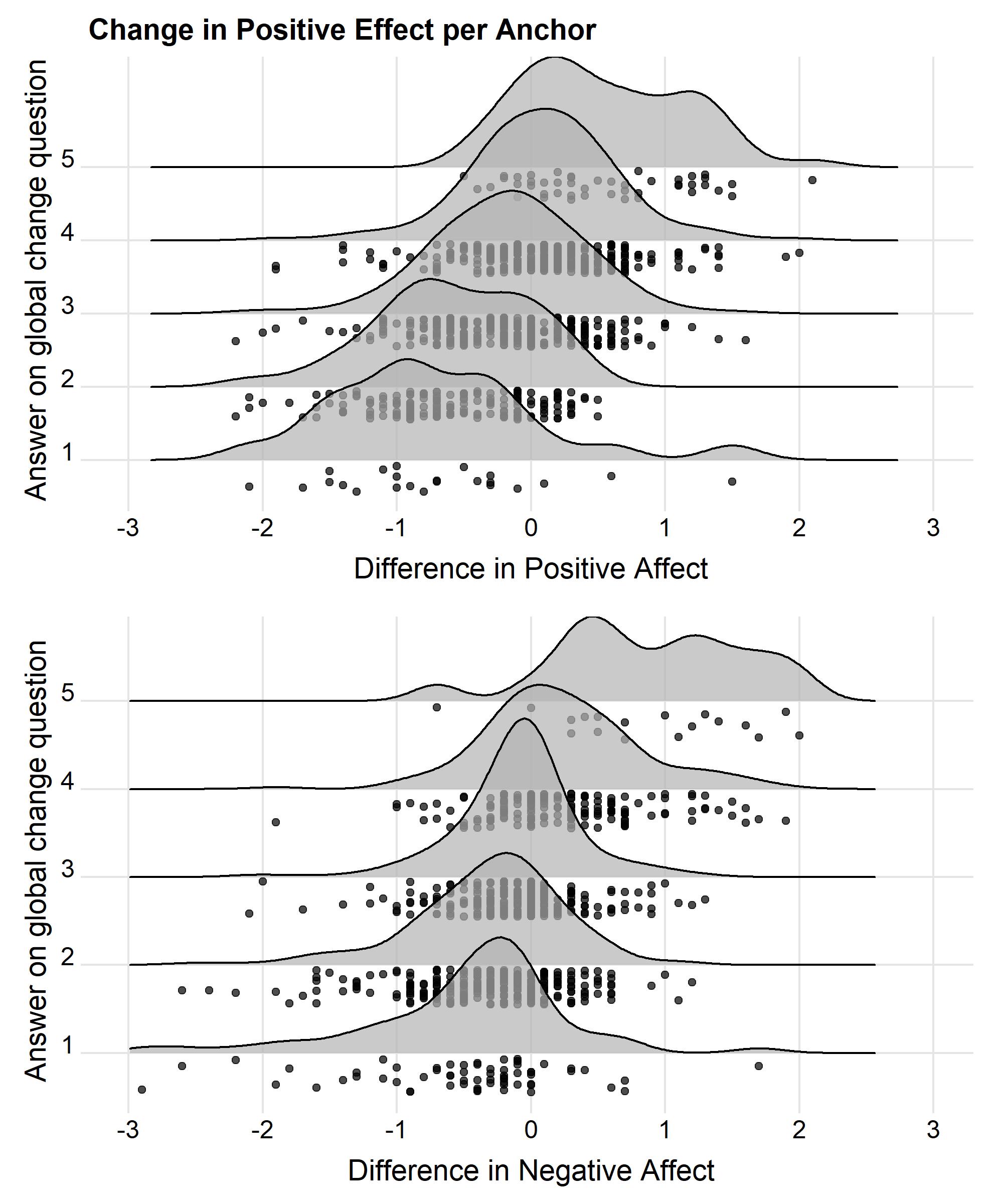
The sampling procedure, methods, and analysis plan for Study 2 were pre-registered on the Open Science Framework (<https://osf.io/a5pze>). The study was identical to Study 1, with the exception that to examine the generalizability of our estimates, the time between T1 and T2 was either 2 days (as in 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. 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](https://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 minimum 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), or as a standardized effect size (e.g., Cohen's *d* of 0.3). Raw scores have the advantage that they do not depend on the standard deviation of the measurements and are recommended for validated scales when researchers always 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 *dz*) or not (Cohen's *dav*). Cohen's *dz* is used in power analyses and in equivalence tests if the equivalence bounds are set in terms of a standardized effect size. Cohen's *dav* 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. We recommend reporting both raw differences and standardized effect sizes.



*Figure 1*. Distribution and individual datapoints of the differences in positive and negative affect as a function of the answer on the global change question.

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), as well as standardized effect sizes and their 95% confidence intervals, subcategorizing participants based on their responses to the GRoC. We focus on participants who selected “a little less positive” or “a little more positive”, and those who selected “a little less negative” or “a little more negative”, for whom the combined results (presented in the “little change” row) are also reported based on a random-effects meta-analysis in which differences for the 'a little less' groups were recoded to be in the positive direction (i.e., multiplied by -1).

First, we see that somewhat surprisingly, participants who report to feel ‘the same’ on the global change question actually show 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 symmetric shifts upward and downward in the ‘a little more/less’ groups for positive and negative affect, this seems to be a general decrease with repeated assessment. Such shifts are observed quite consistently and is referred to as an initial elevation bias in subjective reports (Shrout et al., 2018), although the underlying mechanism is not fully understood. Since the minimum subjectively experienced difference is calculated relative to the ‘no change’ condition, and initial elevation bias impacts all PANAS scores equally over the two time points, the relative differences between the ‘little change’ groups and the ‘no change’ group should still 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 minimum subjectively experienced difference (Hays, Farivar, & Liu, 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 affect.

Table 1

*Combined Dataset: Mean difference [and 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* | Mean Difference | Cohen’s *dz* | Cohen’s *dav* |
| **Positive:** |  |  |  |  |
| MUCH LESS | 23 | -0.71[-1.05;-0.37] | -0.90[-1.38;-0.41] | -0.94[-1.45;-0.43] |
| A LITTLE LESS | 160 | -0.58[-0.67;-0.49] | -0.99[-1.19;-0.80] | -0.82[-0.98;-0.67] |
| THE SAME | 264 | -0.18[-0.25;-0.11] | -0.32[-0.44;-0.19] | -0.22[-0.30;-0.13] |
| A LITTLE MORE | 281 | 0.08[0.02;0.15] | 0.15[0.03;0.27] | 0.11[0.02;0.20] |
| MUCH MORE | 47 | 0.55[0.37;0.72] | 0.91[0.57;1.25] | 0.74[0.46;1.01] |
| **LITTLE CHANGE** | **441** | **0.31[0.26;0.36]** | **0.57[-0.25;1.39]** | **0.46[-0.23;1.16]** |
| **Negative:** |  |  |  |  |
| MUCH LESS | 64 | -0.48[-0.67;-0.29] | -0.63[-0.89;-0.36] | -0.73[-1.04;-0.42] |
| A LITTLE LESS | 263 | -0.32[-0.38;-0.25] | -0.56[-0.69;-0.43] | -0.47[-0.58;-0.36] |
| THE SAME | 270 | -0.11[-0.16;-0.05] | -0.24[-0.36;-0.12] | -0.15[-0.23;-0.08] |
| A LITTLE MORE | 160 | 0.22[0.13;0.31] | 0.38[0.22;0.54] | 0.34[0.20;0.48] |
| MUCH MORE | 18 | 0.88 [0.53;1.24] | 1.24 [0.61;1.85] | 1.03[0.51;1.54] |
| **LITTLE CHANGE** | **423** | **0.25[0.20;0.31]** | **0.48[0.31;0.65]** | **0.42[0.28;0.56]** |

*Note*. Total *N* = 775. The “little change” groups, for both positive and negative affect, are the “little more” and “little less” groups combined. The standardized effect size estimates for the little change groups are combined meta-analytically with bias-correction.

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 as the estimate of how much change is “needed to differ on average for people to stop rating themselves as ‘about the same’ and start rating themselves as ‘a little better’ or ‘a little worse’” (Redelmeier et al., 1993, 1997). In our study, an average difference from the ‘no change’ group of 0.26 (*CI95%* [0.17, 0.36]) scale points led participants to report feeling ‘a little more positive’, and a change of -0.39 (*CI95%* [-0.28, -0.51]) scale points led to feeling ‘a little less positive’. For negative affect, people needed to change 0.20 (*CI95%* [0.12, 0.30]) scale points to stop rating themselves as feeling the same and start rating themselves as feeling a little less negative and they needed to change by -0.36 (*CI95%* [-0.22, -0.43]) scale points to stop rating themselves as feeling the same and start rating themselves as feeling a little more negative.

It is common to combine data from the ‘a little less’ and ‘a little more’ groups into a single estimate of the minimum subjectively experienced difference. 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 GRoC method is applied to. We can see from the 95% CI around the change scores reported above that the effects are asymmetrical. The difference that leads participants to subjectively report feeling ‘a little more’ positive (*M* = 0.26) and ‘a little less’ negative (*M* = 0.20) are statistically smaller than the difference that lead participants to subjectively report feeling ‘a little less’ positive (*M* = -0.39) and ‘a little more’ negative (*M* = -0.36). Although the combined estimates can be computed for both ‘little change’ groups for positive (*M* = 0.31, 95% CI[0.26; 0.36], *SD* = 0.57) and negative (*M* = 0.25, 95% CI[0.20; 0.31], *SD* = 0.57) affect, we believe the observed asymmetry should probably be taken into account when using a minimum subjectively experienced difference in practice. The extent to which such asymmetries occur in other psychological measures is a topic for future research. On a related note, on could consider combining the ‘a little more positive’ and ‘a little less negative’ estimates, and although these estimates are very similar, we hesitate to do so because theoretically positive and negative affect are seen as at least partially independent constructs. Therefore, we believe the four estimates presented above might be the best level of description for researchers who want to specify a SESOI based on the minimum subjectively experienced difference in the PANAS.

The estimate for the minimum subjectively experienced difference is based on the group average. We can therefore expect variability within each group (as visualized in Figure 2), 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 GRoC (e.g., an individual who reports feeling a little more positive at T2 can have a negative change score).

A screenshot of a map

Description automatically generated

*Figure 1.* 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 (GRoC) approach is an anchor-based method that can be used to determine the minimum subjectively experienced difference for various outcome measures in psychology. We provided an illustrative example by estimating the minimum subjectively experienced difference for positive and negative affect as measured by the PANAS. We provide estimates for the relative difference, compared to those participants who report feeling ‘the same’, that lead participants to subjectively report feeling ‘a little more’ positive (*M* = 0.26) and ‘a little less’ negative (*M* = 0.20), as well as the difference that lead participants to subjectively report feeling ‘a little less’ positive (*M* = -0.39) and ‘a little more’ negative (*M* = -0.36). These estimates (or their corresponding standardized effect sizes) can be used as the smallest effect size of interest (SESOI) 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). Our primary recommendation is not to adopt these specific estimates as a smallest effect size of interest, but to adopt anchor-based methods to determine the smallest effect size of interest for the measure you use in the populations you study.

Since the GRoC method only requires adding a single anchor-item (for each domain of interest) to longitudinal designs, collecting such data should be feasible and easily implemented in existing research designs, such as when researchers examine test-retest reliability. By meta-analytically combining estimates, fields can establish precise estimates of the smallest effect size associated with the smallest meaningful 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 points 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; Wright et al., 2012). Because high-precision estimates require large sample sizes (Maxwell, Kelley, & Rausch, 2008) establishing minimum subjectively experienced differences in specific research fields in psychology would benefit from a coordinated approach to data collection.

The GRoC 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. In the Supplemental Materials, we present the results of a between-persons paradigm where people complete the positive items of the PANAS, discuss their current levels of positive affect with a partner, and then complete an anchor question about whether they believe they currently feel the same, slightly more or less positive, or a lot more or less positive, than their interaction partner—the estimates were similar to the within-person approach in the present paper. This between-person approach has been used in past research to estimate the minimum subjectively perceived difference in walking ability (Redelmeier, Bayoumi, Goldstein, & Guyatt, 1997). Such alternative paradigms might be worthwhile to examine in the future.

There are several possible limitations of the usefulness of the GRoC anchor-based method (see, King, 2011; Walters & Brazier, 2003; Wright et al., 2012). First, sometimes eshould Tis primarilyl.

Second, people’s retrospective responses on the GRoC question may more strongly reflect their present state than their change over time (Norman et al., 1997; Wright et al., 2012), while responses on the GRoC question should correlate with their change scores (Cella et al., 2002). In our data, people’s responses on the GRoC question were more strongly correlated with change scores (PA: *r* = .47; NA: *r* = .42) than with T2 present state scores (PA: *r* = .37, *r*dif= 0.10, 95% CI [0.04; 0.16], NA: *r* = .27, *r*dif= 0.14, 95% CI [.07; 0.22]; Zou, 2007). According to Norman et al. (1997), responses on the GRoC should correlate equally with T1 and T2 scores—GRoC ratings 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 GRoC were far less strongly correlated with T1 scores (PA: *r* = -.004; NA: *r* = -.10) than with T2 scores.

There are interesting psychometric questions for future research if researchers adopt anchor-based methods. One question already raised is whether the scores of those who report a little change in the positive and negative direction should be combined into a single estimate, or whether both estimates should be reported separately. 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 minimum subjectively experienced difference separately for those who improve and those who worsen (e.g., Angst, Aeschlimann, & Stucki, 2001). Our results suggest that, for the positive affect dimension of the PANAS, the subgroup of participants who indicated feeling a little more positive had a mean change in positive affect that was considerably lower, in absolute value, than those who said they felt a little less positive. We believe this is also the best approach for changes in affect as measured by the PANAS due to the asymmetry in responses. Additional future questions concern the reliability of the typical single item global change question (versus a multiple item anchor), and exploring context effects and individual differences in the minimum subjectively experienced difference.

To conclude, we believe that the GRoC method holds promise as one possible approach to estimate the minimum 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 study in relation to whether observed effects are large enough to be deemed subjectively meaningful be individuals. In the end, we hope anchor-based methods will help researchers to think more carefully about which effects they consider meaningful.

**Disclosures**

The preregistration for Study 1 can be found at <https://osf.io/b3z65/>.

All data, materials, and code for both studies can be found at <https://osf.io/89pcf/>.

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.

**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.

**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/>).

**References**

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

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

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. https://doi.org/10.1037/1528-3542.6.4.587

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.

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. https://doi.org/10.1037/a0024700

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.*

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

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

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

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.

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

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

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

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

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. (1996). Assessing the Minimal Important Difference in Symptoms: A Comparison of Two Techniques. *Journal of Clinical Epidemiology, 49*(11), 1215-1219. https://doi.org/10.1016/s0895-4356(96)00206-5

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

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>

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. 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), 4. 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). 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

1. Some participants completed the survey one or two days late. [↑](#footnote-ref-1)