Meta-Psychology, 2023, vol 7, MP.2022.3174
https://doi.org/10.15626/MP.2022.3174
Article type: Original article
Published under the CC-BY4.0 license

Open data: Yes
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Matt Williams
Reviewed by: Thom Baguley, Natalie Beribisky
Analysis reproduced by: Lucija Batinović
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/WJMVC

Validity of the Anchor in Estimating the Smallest Subjectively Experienced Difference: Presenting an Anchor-Item Before vs After the Outcome Measure.

Farid Anvari[1]
[1]University of Cologne

In some fields of research, psychologists are interested in effect sizes that are large enough to make a difference to people's subjective experience. Recently, an anchor-based method using a global rating of change was proposed as a way to quantify the smallest subjectively experienced difference—the smallest numerical difference in the outcome measure that, on average, corresponds to reported changes in people's subjective experience. According to the method, the construct of interest is measured on two occasions (Time 1 and Time 2). At Time 2, people also use an anchor-item to report how much they experienced a change in the construct. Participants are then categorized as those who stayed the same, those who changed a lot, and those who changed a little. The average change score for those who changed a little is the estimate of the smallest subjectively experienced difference. In the present study, I examined two aspects of the method's validity. First, I tested whether presenting the anchor-item before or after the Time 2 outcome measure influences the results. The results suggest that any potential influence of the anchor-position, assuming there is an influence, is likely to be small. Second, I examined the anchor-item's validity correlations when the delay between Time 1 and 2 is one day to also see if the pattern is similar to past research where the delay was two and five days. The observed pattern of validity correlations was very similar. I note directions for future research.

Keywords: Smallest effect size of interest, positive affect, negative affect, minimum important difference, subjectively experienced difference, perceptions of change

Introduction

How big should an effect size be for it to be considered theoretically important? Numerous papers have been published on the topic of which effect sizes should be consider "small" or "large" and whether small effect sizes may be important (e.g., Anvari and Lakens, 2021; Bosco et al., 2015; Cafri et al., 2010; Funder and Ozer, 2019; Gignac and Szodorai, 2016; Götz et al., 2021; Hemphill, 2003; Hill et al., 2008; Lovakov and Agadullina, 2021; Nye et al., 2018; Paterson et al., 2016; Plonsky and Oswald, 2014; Richard et al., 2003; Szucs and Ioannidis, 2017; Taylor et al., 2018; Wiernik et al., 2013). However, determining whether an effect size is theoretically important is a more complex task that will depend on the field, measure, and research questions or theoretical claims. In some fields, such as affect and emotion science, many research questions are concerned with people's subjective experiences (e.g., Campbell-Sills et al., 2006; Coutinho and Cangelosi, 2011; Gross, 1999; Kuppens, 2019; LeDoux, 2014; LeDoux and Hofmann, 2018; Reisenzein, 2009; Troy et al., 2018). For such research questions, one boundary between a theoretically important and unimportant effect size could be the smallest difference on the outcome measure that people subjectively experience (Anvari & Lakens, 2021). For example, a study investigating the impact of meditation on mood might consider an effect size important only to the extent that any change in mood is at least large enough for people to notice that difference in their subjective experience such that they would subsequently report that they feel different—anything smaller than this would be too small and thus not important.

Recently, Anvari and Lakens (2021) presented an anchoring method typically used in health research to determine clinically important effect sizes (e.g., Button et al., 2015; Devji et al., 2020; Dworkin et al., 2008; Ebrahim et al., 2017; Guyatt et al., 2002; Jaeschke et al., 1989; King, 2011; Kounali et al., 2020; Norman et al., 2003; Walters and Brazier, 2003), and demonstrated how the anchor-based method can be used to determine the smallest subjectively experienced difference in a measure of mood. I refer readers to Anvari and Lakens (2021) for full details of when and how the anchor-based method can be used and the justifications for the approach. Essentially, the method uses an anchor-item to ask people to give a global rating of change, comparing how they feel now (at Time 2) with how they felt earlier (at Time 1). This global rating of change is then used to categorize people into those who feel a little different (i.e., a little less, or a little more) on the construct of interest, as distinct from those who either feel the same or very different. For those who reported feeling a little different, the average change scores in the outcome measure from Time 1 to Time 2 provides an estimate of the smallest subjectively experienced difference—the smallest difference in the measure that is needed for people to report a difference in how they feel.

To demonstrate the approach, Anvari and Lakens (2021) estimated the smallest subjectively experienced difference in positive and negative affect, as measured by the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988). Specifically, they measured current positive and negative affect at Time 1 and Time 2, separated by either 2 or 5 days. At Time 2, participants also responded to two anchor items asking them how much more/less positive and negative they felt now compared to Time 1 (“much less positive/negative”, “a little less positive/negative”, “about the same”, “a little more positive/negative” or “much more positive/negative). They subcategorized participants into 5 groups for positive affect and 5 groups for negative affect, based on their responses to the respective anchor items. For each subcategory, they then calculated the average change score on the relevant subscale of the PANAS from Time 1 to Time 2. The absolute change score in positive affect ratings for people who either said that they felt a little less or a little more positive provided an estimate of the smallest change in the positive affect subscale of the PANAS that is large enough for people to notice and report. They did the same for the negative affect subscale. Critically, several questions regarding the method’s validity remain unanswered. The purpose of the present paper is to examine two issues regarding the anchor-based method’s validity.

The main question this paper addresses regarding the method’s validity is whether the results vary as a function of presenting the anchor-item before or after the Time 2 measurement. The more that the estimates from this method, and people’s self-reports more generally, vary based on the anchor-item’s position, the more the validity of the method is undermined. This is because such variation would suggest that the anchor’s position matters for people’s self-perceptions of change, leaving researchers with the dilemma of using the “correct” anchor position. In the present study the anchor-item’s position was varied so that it was either presented before or after participants responded to the PANAS at Time 2.

With a single day between measurements, this study also allows me to qualitatively compare the pattern of validity correlations to the pattern found in past research where the time between measurements was 2 and 5 days. Given that the anchor-item relies on people’s memories, the pattern of validity correlations may vary with different time-delays between the two measurements (e.g., a one-day interval). However, because I did not experimentally manipulate the interval between measurements, and nor do I quantitatively compare the estimates from this study with the estimates from previous work, I cannot make causal claims or draw quantitative inferences. The method’s validity is demonstrated by the strength of the correlation between the anchor-item and: (i) the change scores on the outcome measure (e.g., PANAS difference between Time 1 and Time 2), with a positive correlation showing that the anchor-responses capture changes in the construct; (ii) the Time 2 scores on the outcome measure (e.g., PANAS at Time 2), with a positive correlation showing that the anchor-responses reflect how people currently feel; and (iii) the Time 1 scores on the outcome measure (e.g., PANAS at Time 1), with a negative correlation showing that the anchor-responses also reflect people’s feelings at the time against which current feelings are being compared (see Devji et al., 2020; Kamper et al., 2009). For excellent validity, (i) should be stronger than (ii) and (iii), showing that the anchor-responses capture change in feelings more strongly than current or past states. Moreover, (ii) and (iii) should be of equal magnitude, or as close to as possible, though in opposing directions, showing that the anchor-responses reflect current feelings and past feelings to the same extent.

Anvari and Lakens (2021) found that although the anchor-item showed some validity, there were issues that needed further examination. They used two- and five-day intervals between measurements and found that although responses on the anchor-item were more strongly correlated with change scores than with Time 2 and Time 1 scores, the correlations with the Time 2 scores were much stronger than the correlations with Time 1 scores—the latter correlations were almost zero for positive affect. This suggested that people’s judgments of change reflected their present Time 2 state too strongly, relative to their past Time 1 state. Anvari and Lakens (2021) argued that these validity issues may have been caused by memory biases. Indeed, there is ample evidence that memory for past feelings as measured by self-report rating scales is inaccurate (e.g., Levine et al., 2018; Levine et al., 2009; Robinson and Clore, 2002). because the present study examines the validity correlations with an interval of one day between the Time 1 and Time 2 measures, I can see if the general pattern of validity correlations is similar to past research, or if it is better (e.g., stronger correlation of anchor responses with Time 1 measures). Once again, however, because I did not manipulate the time-delay, I can make no causal or quantitative inferences about the time between measurements.

In this study, I aimed to examine whether, and by how much, the results would be altered by presenting the anchor item before vs after the measure of interest. Specifically, I wanted to see whether presenting the anchor before vs after the measure of interest would (i) change the estimate of the smallest subjectively experienced difference; (ii) change people's ratings on the measure of interest; (iii) whether a time-delay of 1-day between measurements would show a similar pattern of the anchor item's validity correlations as compared to past research that used 2- or 5-days; and (iv) whether the correlations in (iii) varied as a function of the anchor item's position.

Methods

An important point that I'd like to note is that the estimates of the smallest subjectively experienced difference in this study should not be used as the basis for a smallest effect size of interest in other studies or, otherwise, used with caveats such as that the smallest subjectively experienced difference might vary based on the population of interest. The purpose of this study was to examine factors that might affect the estimates that the method produces. To produce reliable and valid estimates that are useable, I believe that a lot more work should be done.

The Supplemental Materials, data and analysis code for this paper can be found on the Open Science Framework (https://osf.io/vjx6c/).

Procedure

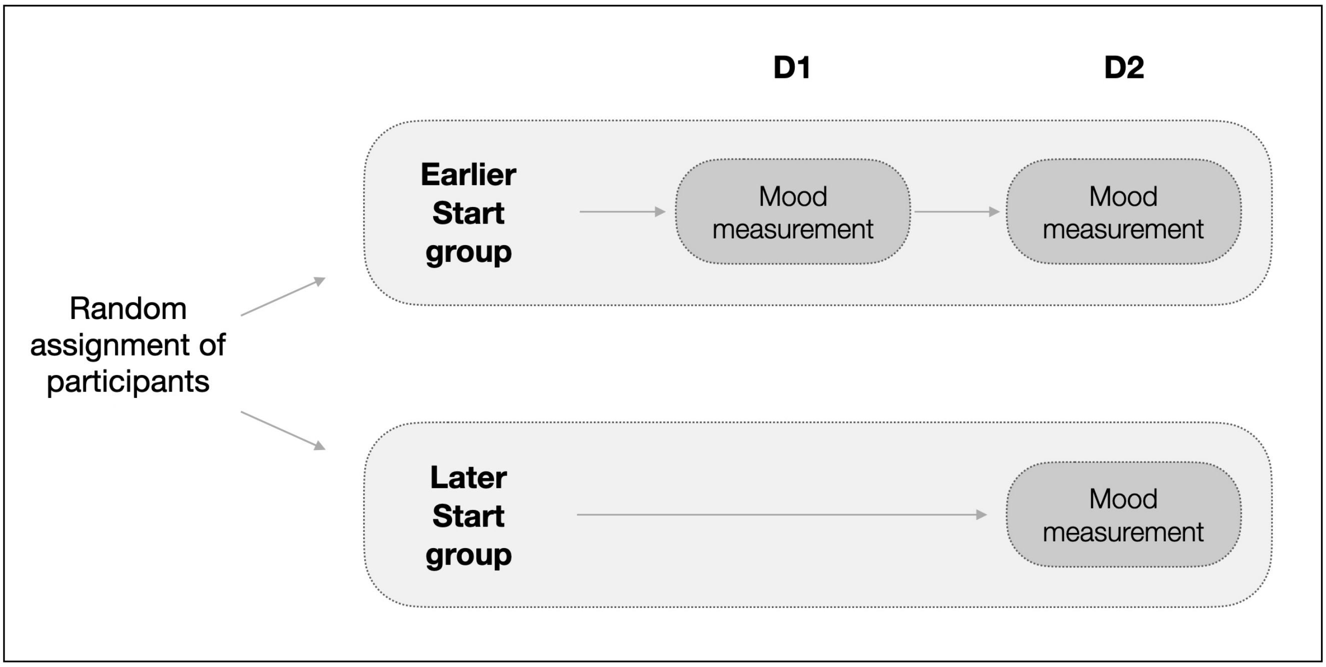
I included the anchor-item in a study designed to address an unrelated research question on the robustness of the initial elevation bias in measures of mood (Shrout et al., 2018). For that study, I recruited 2,306 participants from Prolific.co, requiring only that they be fluent in English. Some participants responded to the PANAS items on two measurement occasions separated by one day, which provided an opportunity to include an anchor-item for addressing the research questions in the present paper. The design of the full study is presented in Figure 1 and the explanation follows. Participants were recruited on a Monday and randomly allocated to either the Earlier Start group who would take the survey twice (N = 1,150) or to the Later Start group who would take the survey once (N = 1,156). The present paper focuses only on the Earlier Start group for which there were two measurement occasions. The participants in the Earlier Start group were invited to participate in the study on the following day (i.e., Tuesday) at which point they completed a 3-item anxiety scale and a 3-item vigour scale (Cranford et al., 2006), followed by the 20-item PANAS. Only the PANAS is relevant for the present paper. This first survey provided the Time 1 measure. There were 1,011 participants who completed the Time 1 measure, after removing 7 duplicate entries, and who were subsequently invited to complete the survey again one day later (i.e., on Wednesday). On Wednesday, all participants were invited to take the exact same survey involving the same measures in the same order. This second survey provided the Time 2 measure. In the survey at Time 2, I included an anchor-item for positive affect and an anchor-item for negative affect. Importantly, participants were randomly allocated to respond to the anchor item either before the PANAS or after.

Participants

Power calculations were not performed for the present paper because the original study was designed for another purpose. Nevertheless, given that there were 903 participants in the Earlier Start group, whose data are used for the present study, compared to the 775 participants in Anvari and Lakens (2021), the current data are likely to provide reliable and informative results for examining the pattern of validity correlations and the possible impact of the anchor-items' positions. I excluded participants for (i) having missing responses on any measures or (ii) giving the same response to all items across the three measures in either the first or second survey. The participants with missing responses had no data for any of the PANAS items at all, and giving the same response to all items indicates inattentive or unthoughtful responses. The final dataset for this paper consisted of 903 participants who completed the PANAS on both Time 1 and Time 2. Of these, there were 440 participants who responded to the anchor-items before the PANAS at Time 2 and 463 who responded to the anchor-items after the PANAS. Of the 903 participants who completed both Time 1 and Time 2 measures, 436 identified as female, 454 identified as male, 2 identified as trans male, 8 identified as gender queer or nonconforming, and 2 had not revealed their gender identity. Participants had a mean age of 26.9 years (SD = 8.3), ranging from 18 years to 71 years. Participants resided in 31 different countries (a detailed list of the number

Figure 1

Schematic diagram of the original study design.



Note. Participants were recruited and randomly allocated to two groups, the Earlier Start group and the Later Start group. The day after recruitment, D1, participants in the Earlier Start group completed the mood measures, including the PANAS. The day after that, D2, participants in the Earlier Start group (and those in the Later Start group) completed the mood measures, including the PANAS, and responded to the anchor-items. Only participants in the Earlier Start group are relevant for the present paper.

of participants from each country is on the first page of the Supplemental).

Measures

The outcome measure of interest were ratings on the PANAS, which participants completed at both Time 1 and Time 2. Participants were told that, “This scale consists of a number of words that describe different feelings and emotions. Read each item and then indicate how much you feel this way right now.”. Participants could then rate each of the emotions, presented in a matrix in random order, on 5-point Likert scales (from 1 = 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). For Time 1 and Time 2 separately, I averaged the positive affect items for each participant to produce a mean score for positive affect, and I did the same for the negative affect items to produce a mean score for negative affect. To calculate each participant’s change score for positive affect, I subtracted the mean score for positive affect at Time 1 from the mean score for positive affect at Time 2. I did the same to calculate the change scores for negative affect. Thus, each participant had a change score for positive affect and a change score for negative affect.

The anchor items for positive and negative affect included in the survey at Time 2 asked participants, “Overall, compared to the time when you did this survey yesterday, how positive/negative do you feel right now?”. Participants had five response options for each anchor item: much less positive/negative, a little less positive/negative, the same, a little more positive/negative, and much more positive/negative.

Results & Discussion

In the analyses I used Welch’s independent samples *t*-tests. As per the anchor-based approach for determining the smallest subjectively experienced difference in affect (Anvari & Lakens, 2021), I subcategorized participants into five groups based on their responses to the anchor item and, for each subcategory, I calculated the average change score from Time 1 to Time 2. I did this separately for participants who saw the anchor-item before the PANAS and those who saw the anchor-item after. The results for positive and negative affect are presented in Tables 1 and 2, respectively, which show the Time 1 scores, Time 2 scores, and the average change scores (the Tables in the Supplemental Materials also contain the mean differences standardized such that they either take into account the correlation between the measures at Time and Time 2 (labeled Cohen's dz) or not (labeled Cohen's dav)—the Supplemental Materials also has a table that provides the estimates for the combined before and after groups). The formula for Cohen's dz it is:

m_diff / s_diff, where m_diff is the mean difference between the T1 and T2 measurements, and s_diff is the standard deviation of the mean difference.

And for Cohen's dav the formula is:

m_diff / ((sd1+sd2)/2), where the denominator is the averaged standard deviation of T1 and T2 measurements—i.e., sd1 and sd2 are the standard deviations of T1 and T2 measurements, respectively.

The first aim was to examine the extent to which the estimates, and people's self-reports more generally, vary as a function of whether the anchor-item is presented to participants before or after the Time 2 outcome measure. For the estimates of the smallest subjectively experienced difference, the relevant subcategories of participants from Tables 1 and 2 are those who said they felt "a little less" positive/negative and those who felt "a little more" positive/negative.

For positive affect, the average change scores for people who said they felt a little less positive converged almost perfectly, regardless of whether people saw the anchor before (*M* = -0.54, *SD* = 0.68) the PANAS or after (*M* = -0.56, *SD* = 0.69), mean difference = 0.02, CI 95%[-0.21, 0.26], *d* = 0.03, CI 95%[-0.31, 0.38]; although the confidence intervals are wide due to the relatively small number of people who said that they felt "a little less positive". For those who said they felt a little more positive, there was a slight difference in the average change score between those who saw the anchor before (*M* = 0.20, *SD* = 0.53) and those who saw the anchor after (*M* = 0.15, *SD* = 0.63), but this was not a statistically significant difference, *t*(351.87) = 0.83, *p* = .409, mean difference = 0.05, CI 95%[-0.07, 0.17], *d* = 0.09, CI 95%[-0.12, 0.29]. The observed effect size was quite tiny. Moreover, the average change score for people who said that they felt "the same" was negative, regardless of whether they saw the anchor before or after the PANAS; and for these "same" groups, the results also converged almost perfectly regardless of the anchor-position (*M*before = -0.14, *SD*before = 0.60; *M*after = -0.16, *SD*after = 0.53), mean difference = 0.02, CI 95%[-0.12, 0.16]. These combined results suggest that for positive affect the estimates had very little, if any, variation as a function of the anchor-item's position. Hence, the estimates of the smallest subjectively experienced difference, derived using the anchor-item, are unlikely to be strongly impacted by whether it is presented before or after the Time 2 measure of positive affective states.

For negative affect, of the people who said they felt a little less negative, those who saw the anchor before (*M* = -0.22, *SD* = 0.50) the PANAS had a slightly lower change score, in absolute terms, as compared to those who saw the anchor after (*M* = -0.29, *SD* = 0.56), though this difference wasn't statistically significant, *t*(347.49) = 1.10, *p* = .272, mean difference = 0.07, CI 95%[-0.05, 0.17], *d* = 0.12, CI 95%[-0.09, 0.33]. As for positive affect, the raw and standardized effect size estimates were quite small. For people who said they felt a little more negative, those who saw the anchor before (*M* = 0.41, *SD* = 0.58) the PANAS had a slightly higher change score as compared to those who saw the anchor after (*M* = 0.26, *SD* = 0.67), though this difference was also not statistically significant, *t*(133.53) = 1.40, *p* = .165, mean difference = 0.15, CI 95%[-0.03, 0.36], *d* = 0.24, CI 95%[-0.10, 0.58]. The effect sizes here are a little larger than for the previous comparisons and the wider confidence intervals include relatively large effect sizes, likely due to the small samples in these groups. Moreover, people who said that they felt "the same" in negative affect showed an average change score that was in the negative direction regardless of whether they saw the anchor before or after the PANAS. However, of these participants, those who saw the anchor before (*M* = -0.09, *SD* = 0.50) the PANAS had a slightly larger absolute change score than those who saw the anchor after (*M* = -0.02, *SD* = 0.47), though this difference was small and not statistically significant, *t*(235.93) = 1.12, *p* = .262, mean difference = 0.07, CI 95%[-0.05, 0.19], *d* = 0.15,, CI 95%[-0.11, 0.40]. Although these results provide suggestive evidence that the anchor position may influence people's self-perceptions of change for negative affect, strong conclusions should not be drawn due to the relatively few people who reported that they felt "a little more negative" in negative affect. Indeed, the estimates for people who said they felt "a little more" negative have wider confidence intervals relative to the confidence intervals for those who said they felt "a little less" negative, very likely because the former has more than double the sample size.

To further examine any potential differences between

Table 1

Positive Affect: Means (Standard Deviations) and Average Change Score [95% Confidence Intervals] in PANAS scores from Time 1 to Time 2, with participants subcategorized based on their responses to the anchor-item.

	N	T1: M (SD)	T2: M (SD)	Mean Change Score
Before				
Much less	19	2.43 (0.87)	1.71 (0.46)	-0.72 [-1.02, -0.43]
A little less	62	2.94 (0.82)	2.40 (0.72)	-0.54 [-0.71, -0.37]
The same	120	2.83 (0.81)	2.69 (0.83)	-0.14 [-0.25, -0.03]
A little more	187	2.79 (0.82)	2.99 (0.85)	0.20 [0.13, 0.28]
Much more	52	2.98 (1.08)	3.60 (0.91)	0.62 [0.42, 0.83]
After				
Much less	26	2.87 (0.73)	1.81 (0.41)	-1.06 [-1.38, -0.74]
A little less	71	2.92 (0.82)	2.35 (0.75)	-0.56 [-0.73, -0.40]
The same	128	2.81 (0.85)	2.65 (0.86)	-0.16 [-0.25, -0.06]
A little more	181	2.73 (0.91)	2.88 (0.82)	0.15 [0.06, 0.25]
Much more	57	2.98 (1.11)	3.72 (0.77)	0.74 [0.53, 0.96]

Note. Total *N* = 903. Change Score = Average Change Score. Before = anchor-item presented before the PANAS. After = anchor-item presented after the PANAS. 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.

Table 2

Negative Affect: Means (Standard Deviations) and Average Change Score [95% Confidence Intervals] in PANAS scores from Time 1 to Time 2, with participants subcategorized based on their responses to the anchor-item.

	N	T1: M (SD)	T2: M (SD)	Mean Change Score
Before				
Much less	71	1.94 (0.79)	1.40 (0.52)	-0.54 [-0.72, -0.37]
A little less	169	1.88 (0.70)	1.66 (0.63)	-0.22 [-0.30, -0.15]
The same	123	1.88 (0.79)	1.80 (0.71)	-0.09 [-0.18, 0.00]
A little more	67	1.95 (0.79)	2.37 (0.83)	0.41 [0.27, 0.55]
Much more	10	2.79 (1.15)	3.67 (0.94)	0.88 [0.26, 1.50]
After				
Much less	72	1.71 (0.61)	1.36 (0.49)	-0.35 [-0.52, -0.18]
A little less	181	1.92 (0.79)	1.64 (0.60)	-0.29 [-0.37, -0.20]
The same	115	1.76 (0.74)	1.75 (0.69)	-0.02 [-0.10, 0.07]
A little more	70	2.09 (0.93)	2.35 (0.83)	0.26 [0.10, 0.42]
Much more	25	2.04 (0.79)	3.01 (0.77)	0.96 [0.58, 1.35]

Note. Total N = 903. Change Score = Average Change Score. Before = anchor-item presented before the PANAS. After = anchor-item presented after the PANAS. 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.

people who saw the anchor before vs after the Time 2 measurement, I also conducted *t*-tests on the Time 2 ratings and on the change scores, without subcategorizing participants based on their anchor-responses. These analyses provide two benefits beyond the analyses reported in the preceding paragraphs. First, the results speak more generally to the question of whether people's self-reports and self-perceptions of change vary as a function of the anchor-item's position. This is because the following analyses examine the full sample, rather than focusing only on the "a little less/more" groups. Second, by examining the full sample the statistical tests have higher statistical power.

For positive affect, people who saw the anchor-items before the PANAS (M = 2.84, SD = 0.91) had statistically nonsignificant differences in their Time 2 ratings compared to people who saw the anchor-items after (M = 2.78, SD = 0.91), t(898.27) = 1.03, p = .305, mean difference = 0.06, CI 95%[-0.06, 0.18], d = 0.07, CI 95%[-0.06, 0.20]. Similarly, change scores in positive affect for people who saw the anchor items before the PANAS (M = 0.02, SD = 0.70) were not statistically significantly different from change scores for people who saw the anchor items after (M = -0.04, SD = 0.78), t(898.40) = 1.07, p = .283, mean difference = 0.02, CI 95%[-0.04, 0.15], d = 0.07, CI 95%[-0.06, 0.20]. In both cases, the observed raw and standardized effect sizes are quite tiny. For negative affect, people who saw the anchor-items before (M = 1.81, SD = 0.79) had statistically nonsignificant differences in Time 2 scores compared to people who saw the anchor-items after (M = 1.80, SD = 0.77), t(896.97) = 0.11, p = .910, mean difference = 0.01, CI 95%[-0.10, 0.11], d = 0.01, CI 95%[-0.12, 0.14]—this difference was centred almost exactly on zero. And the difference in change scores for negative affect between people who saw the anchor items before (M = -0.12, SD = 0.65) the PANAS as compared to after (M = -0.08, SD = 0.69) were also statistically nonsignificant, t(900.89) = 0.83, p = .407, mean difference = 0.04, CI 95%[-0.05, 0.12], d = 0.06, CI 95%[-0.08, 0.19]. Again, the observed raw and standardized effect sizes were tiny.

Given that the tests reported in the preceding paragraph involved the full sample of participants, providing higher statistical power to detect potential differences that might exist, and producing more results, the findings lend additional support to the idea that the anchor-position may not matter too much. Indeed, all of the observed raw and standardized effect sizes were tiny. Moreover, based on the above confidence intervals, if the anchor-position does affect the change scores (and thus the smallest subjectively experience difference) or people's ratings at Time 2, then the effect size for how much it matters is likely to be smaller than d = 0.20. Taken together, this study did not show strong and reliable evidence that presenting the anchor-item before or after the outcome of interest substantially changes people's self-reports or their perceptions of change. On the other hand, the observed effect sizes suggest that the anchor-position should not matter by too much. Nevertheless, it should be easy for researchers collecting longitudinal data with mood measures to include an anchor-item and vary whether it's presented before or after the Time 2 measure. As more data is gathered in this way, and made publicly available, we can meta-analyse the results and potentially rule out even smaller effect sizes, being more confident in the conclusions.

The second aim of this study was to examine the anchor-item's validity correlations and see if a similar pattern emerged as compared to validity correlations reported with longer delays of two to five days (i.e., Anvari and Lakens, 2021). Firstly, for the anchor to be valid, it needs to correlate relatively strongly with the change scores. Indeed, this was the case for both positive affect (r = .54„ CI 95%[.49, .58]) and negative affect (r = .45„ CI 95%[.40, .50]). Although some researchers have suggested that a minimum correlation of .50 is necessary for good validity (Devji et al., 2020), we should be wary of interpreting benchmarks too rigidly. The present results suggest that the anchor-responses are capturing change in the construct.

Secondly, the anchor needs to correlate with the Time 2 scores, which was also the case for both positive (r = .45„ CI 95%[.40, .50]) and negative affect (r = .48„ CI 95%[.42, .52]). Importantly, however, to show that the anchor-responses reflect change in the construct more than present state, the anchor's correlation with the change scores should be stronger than its correlation with Time 2 scores. This was the case for positive affect ($r$dif = .09, CI 95%[.03, .15]) but not for negative affect ($r$dif = -.02„ CI 95%[-.08, 0.04]; Zou, 2007). Critically, the anchor-responses should not reflect the present state more than they reflect change in the construct, and indeed the results show that this was not the case. Hence, the current results are not quite ideal because the anchor item ratings are not more strongly related to change scores than Time 2 for negative affect. At the same time the anchor item ratings are not more strongly related to Time 2 scores than to change scores which would be a major problem. Thirdly, though rarely achieved in practice (Devji et al., 2020), the anchor should correlate with Time 1 scores just as strongly, in absolute terms, as with Time 2 scores. In the present study, much like the results reported by Anvari and Lakens (2021), the anchor's correlations with Time 1 scores were much weaker for both positive affect (r = .01„

CI 95%[-.05, .08]) and negative affect (r = .09,, CI 95%[.02, .15]). Hence, people's anchor-responses reflected their current state much more than their past Time 1 state.

An important consideration for interpreting the above pairwise correlations is that the Time 1 and Time 2 ratings are relatively strongly correlated (rPA = .66, rNA = .63, ps < .001). This can be accounted for by examining the relationship of the anchor responses with the Time 1 and Time 2 ratings in a multiple regression analysis. I therefore regressed the anchor responses on to the Time 1 and Time 2 scores. For positive affect, Time 1 (b = -0.59, p < .001) and Time 2 (b = 0.88, p < .001) scores showed the expected relationship with the anchor responses. Similarly for negative affect, Time 1 (b = -0.48, p < .001) and Time 2 (b = -0.94, p < .001). The 'b' values are unstandardized regression coefficients. Therefore, when the correlation between the current state and their Time 1 is accounted for in regression analyses, the relationship of the anchor responses with the present and past states gets closer to the ideal of equal in magnitude but opposite in direction.

I also tested whether any of the validity correlations varied as a function of the anchor-item's position. I did this by using linear regression analyses in which I, in turn, regressed the change scores, Time 2 scores, and Time 1 scores onto (i) the responses to the anchor-item, (ii) the anchor-item's position, and (iii) the interaction between the latter two. For both positive affect and negative affect, the interactions were all statistically non-significant (all ps > .179). For descriptive purposes, Table 3 presents the correlations of the anchor with the (i) change scores, (ii) Time 2 scores, and (iii) Time 1 scores, separately for participants who saw the anchor before vs after the PANAS. There was little evidence to suggest that the validity correlations reliably varied as a function of whether the anchor-items were presented before or after the Time 2 measure.

Therefore, the pattern of validity correlations for the anchor-item using a one-day interval between measurements was qualitatively similar to the pattern presented in past research using two- and five-day intervals (i.e., Anvari and Lakens, 2021). In both cases, the anchor responses correlated more strongly with change scores than with Time 2 and Time 1 scores, but the correlations with Time 2 scores were much stronger than the correlations with Time 1 scores. Future research can examine whether the validity correlations can be causally improved with experimental tests in which the time interval between measurements is manipulated.

Finally, it's worth noting that some of the other results in the present study had a similar pattern to results reported in Anvari and Lakens (2021). Specifically, for positive affect (see Table 1), the average change score for people who said that they felt "a little less" positive is larger in absolute magnitude than the average change score for people who said that they felt "a little more" positive, regardless of whether they saw the anchor-item before or after the PANAS. Anvari and Lakens (2021) found the same pattern for both positive affect and negative affect, though for negative affect the pattern was less pronounced. In contrast, for negative affect in the present study (see Table 2), the above pattern is reversed for people who saw the anchor-item before the PANAS but not for those who saw the anchor-item after. Moreover, and consistent with past findings (Anvari & Lakens, 2021), for both positive and negative affect, the group of people who said that they felt "the same" showed an average change score in the negative direction. These consistent patterns seem worth exploring in future research.

Conclusion

In sum, the results of the present study provide preliminary answers to some validity questions and suggest directions for future work. First, the anchor-item's position is unlikely to alter the estimates of the smallest subjectively experienced difference, or people's self-reports more generally, by very much, at least for the PANAS, and especially for positive affect. Research should examine (i) whether these findings are generalizable to other measures of affect, or to measures of other constructs (e.g., life satisfaction), and (ii) whether the findings in my study hold for the groups that felt "a little less positive" or "a little more negative", since in my study these groups had small sample sizes and thus wide confidence intervals. Second, the pattern of validity correlations in the present study, with a one-day interval between measures of current mood, was similar to the pattern of validity correlations in studies with two- or five-day intervals between measures of feelings reported for the whole of a day (i.e., retrospective day reports). Future research can examine whether the pattern of validity correlations can be improved by manipulating the interval between measurements for a direct causal quantitative comparison.

One limitation of the present work is that I had no smallest effect size of interest myself. Future research could address this by either defining and preregistering a smallest effect size of interest for the analyses, or by using an extremely large sample size in order to obtain precise estimates of all of the effects. In the latter case, we can then at least know how large the relevant effect sizes are and thus make adjustments to the methods and/or estimates of the smallest subjectively experienced difference accordingly. For example, if a study

Table 3

Correlations [and 95% Confidence Intervals] of the Anchor Responses with Change Scores, Time 2 Scores, Time 1 Scores

	Measures		
	Change Scores	Time 2 Scores	Time 1 Scores
Positive Affect			
Anchor before	.52 [.44, .58]	.44 [.36, .51]	.04 [-.05, .14]
Anchor after	.56 [.49, .62]	.46 [.39, .53]	-.02 [-.11, .08]
Negative Affect			
Anchor before	.47 [.39, .54]	.46 [.38, .53]	.07 [-.02, .16]
Anchor after	.44 [.36, .51]	.50 [.42, .56]	.10 [.01, .19]

found that presenting the anchor before (vs after) the Time 2 measures changed the validity correlations by even a tiny bit then we could use the approach with better validity correlations. Or if a study found that the smallest subjectively experienced difference estimates varied by a small amount depending on the position of the anchor item, then researchers could either use the average of the two, or use the estimate from the method with the better validity correlations.

Importantly, other assumptions of the anchor item remain untested (see Anvari and Lakens, 2021). The most fundamental assumption of the method, and of the whole enterprise of determining the smallest subjectively experience difference, is that it is possible to obtain. Specifically, it is assumed that a smallest subjectively experience difference in the construct can be obtained in a way that is not a result of some methodological artifact. Or, that there is such a thing as "the smallest subjectively experienced difference" in the construct of interest. If we hold that fundamental assumption, there are then other assumptions of the anchor-method that are perhaps more readily testable. For example, it is assumed that people can accurately recall how they felt in the past, report how they feel in the present, and compare their past and current feelings to accurately report a difference on the anchor-item. The validity correlations reported above attempt to address these assumptions. There is, moreover, the assumption that the smallest subjectively experienced difference will be the same regardless of various arbitrary methodological decisions. For example, it is unknown whether and how much the estimate of the smallest subjectively experienced difference changes depending on the number of scale points the anchor item has (e.g., 5- vs 7- vs 11-points). One study with a relatively small sample (N = 181) suggests there may be small differences between anchor items with 7- and 15-points (Lauridsen et al., 2007). Much work is required to assess all of the assumptions of the anchor method. Finally, Importantly, the anchor method is useful only for researchers with an interest in effects that people subjectively experience or consider meaningful. But this will not always be the case. For example, researchers may base their smallest effect size of interest on some sort of cost-benefit analysis or on a theoretically predicted effect size, and these may be smaller than the smallest subjectively experienced difference.

Further work assessing the validity and assumptions of the anchor method are important because this method provides one potential way for researchers to determine the smallest effect size of interest, at least for research questions that are to do with people's subjective experiences. Being able to determine the smallest effect size of interest will be a great advance, as it would allow for researchers to make falsifiable hypotheses, do informative power calculations, and rule out effect sizes that are too small to matter for the research question at hand.

Author Contact

Farid Anvari, University of Cologne, email: faridanvari.fa@gmail.com. https://orcid.org/0000-0002-5806-5654

I would like to thank Daniël Lakens and Iris Schnedier for providing feedback on a draft of this paper.

Conflict of Interest and Funding

There are no conflicts of interest to declare. Farid Anvari was supported by This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 883785, and a European Association of Social Psychology (EASP) seedcorn research grant which provided funds for the data collected.

Author Contributions

I conducted all of the research, draft writing, editing, study design, data collection, and analyses.

Open Science Practices

This article earned the Open Data and the Open Materials badge for making the data and materials openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, are published in the online supplement.

References

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, 104159. https://doi.org/10.1016/j.jesp.2021.104159

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*, 431–449. https://doi.org/10.1037/a0038047

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

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2010). A meta-meta-analysis: Empirical review of statistical power, type i error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research*, *45*(2), 239–270. https://doi.org/10.1080/00273171003680187

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

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

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, *32*(7), 917–929. https://doi.org/10.1177/0146167206287721

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

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

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*, 156–168. https://doi.org/10.1177/2515245919847202

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2021). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691620984483

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

Guyatt, G. H., Osoba, D., Wu, A. W., Wyrwich, K. W., Norman, G. R., & Group, C. S. C. M. (2002). Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings*, *77*(4), 371–383. https://doi.org/10.4065/77.4.371

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, *58*(1), 78–79. https://doi.org/10.1037/0003-066x.58.1.78

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

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

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

Lauridsen, H. H., Hartvigsen, J., Korsholm, L., Grunnet-Nilsson, N., & Manniche, C. (2007). Choice of external criteria in back pain research: Does it matter? recommendations based on analysis of responsiveness. *Pain*, *131*(1), 112–120. https://doi.org/10.1016/j.pain.2006.12.023

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, L. J., Lench, H. C., Karnaze, M. M., & Carlson, S. J. (2018). Bias in predicted and remembered emotion. *Current Opinion in Behavioral Sciences*, *19*, 73–77. https://doi.org/10.1016/j.cobeha.2017.10.008

Levine, L. J., Lench, H. C., & Safer, M. A. (2009). Functions of remembering and misremembering emotion. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *23*(8), 1059–1075. https://doi.org/10.1002/acp.1610

Lovakov, A., & Agadullina, E. (2021). Empirically derived guidelines for interpreting effect size in social psychology [online first publication]. *European Journal of Social Psychology*. https://doi.org/10.1002/ejsp.2752

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

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2018). How big are my effects? examining the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research Methods*, 1094428118761122. https://doi.org/10.1177/1094428118761122

Paterson, T. A., Harms, P. D., Steel, P., & Credé, M. (2016). An assessment of the magnitude of effect sizes: Evidence from 30 years of meta-analysis in management. *Journal of Leadership & Organizational Studies*, *23*(1), 66–81. https://doi.org/10.1177/1548051815614321

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? interpreting effect sizes in l2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

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

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, *128*(6), 934. https://doi.org/10.1037/0033-2909.128.6.934

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

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Taylor, J. A., Kowalski, S. M., Polanin, J. R., Askinas, K., Stuhlsatz, M. A. M., Wilson, C. D., Tipton, E., & Wilson, S. J. (2018). Investigating science education effect sizes: Implications for power analyses and programmatic decisions. *AERA Open*, *4*(3). https://journals.sagepub.com/doi/10.1177/2332858418791991

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

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

Wiernik, B. M., Ones, D. S., & Dilchert, S. (2013). Age and environmental sustainability: A meta-analysis. *Journal of Managerial Psychology*, *28*(7/8), 826–856. https://doi.org/10.1108/jmp-07-2013-0221

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