Bias in Self-Reports: An Initial Elevation Phenomenon

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Bias in Self-Reports: An Initial Elevation Phenomenon

Farid Anvari[1] [ID]**, Emir Efendić[2], Jerome Olsen[3],
Ruben C. Arslan[4,5], Malte Elson[6], and Iris K. Schneider[1]** [ID]

Abstract

Researchers have long worried about a phenomenon where study participants give higher ratings on self-report scales the first time they take a survey compared to subsequent times, particularly for negative subjective experiences. Recent experimental evidence, using samples of U.S. college students, suggests that this initial elevation phenomenon is due to an upward bias in people's initial responses. Such bias potentially undermines the validity of many research findings. However, more recent studies have found little evidence in support of the phenomenon. To investigate the robustness of the initial elevation phenomenon, we conducted the largest experiments to date in diverse online samples (N = 5,285 across three studies, from Prolific.co). We observed an initial elevation on self-reports of negative subjective experiences such as mood and mental and physical health symptoms. Our findings show that the threats to validity posed by the phenomenon are real and need to be reckoned with.

Introduction

Various subfields of psychology have increasingly incorporated affect, mood, and emotions into their models of human behavior and mind (Dukes et al., 2021). Arguably, the most common way of measuring affect, mood, and emotions is with self-report rating scales. For example, people might report how intensely they feel an emotion, from 1 = *not at all* to 5 = *extremely*.

Despite the overwhelming popularity of rating scales, a question of validity has troubled researchers for over half a century. For a wide range of affect and personality measures, participants seem to give higher ratings the first time they take a survey compared to subsequent times, particularly for negative feelings such as anxiety and depression (e.g., Brantley et al., 1988; French & Sutton, 2010; Ganzach & Bulmash, 2021; Knowles et al., 1996; Lucas et al., 1999; Milich et al., 1980; Neprash, 1936; Piacentini et al., 1999; Reynolds et al., 2016; Ribera et al., 1996; Sharpe & Gilbert, 1998; Windle, 1954, 1955).

Until recently, this initial elevation phenomenon could be explained away as a problem with study design. For example, the phenomenon could be caused by the confounds of time, with the decrease in ratings being a result of changes in people's feelings due to some event occurring between the first and subsequent reports. Alternatively, the phenomenon could be due to sampling bias, where people higher on the attribute being measured (e.g., anxiety) would be selected into the study and (1) subsequently feel better with time or

(2) drop out of the study leaving the less anxious people in the sample for the second measurement occasion (Arslan et al., 2021; Iachina & Bilenberg, 2012; Milich et al., 1980).

However, recent *experimental* investigations suggest a deeper and more concerning problem; namely, the initial elevation phenomenon is an upward *bias* in people's first response on rating scales (Shrout et al., 2018). Across four studies using U.S. college samples, Shrout et al. (2018) randomly assigned participants to groups who would start a longitudinal study at different, but overlapping, times. The researchers could therefore compare two groups' ratings on the same scale given on the same day, with the only difference being that one group had responded to the scales before. Thus, the confounds of time and sampling bias would have affected both groups similarly so that these could not be the cause of any differences. The results showed that mean ratings were consistently higher for people responding to the scales for the first time as compared

[1]University of Cologne, Germany
[2]Maastricht University, The Netherlands
[3]Max Planck Institute for Research on Collective Goods, Bonn, Germany
[4]University of Leipzig, Germany
[5]Max Planck Institute for Human Development, Berlin, Germany
[6]Ruhr University Bochum, Germany

Corresponding Author:
Farid Anvari, Department of Psychology, Social Cognition Center Cologne, University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany.
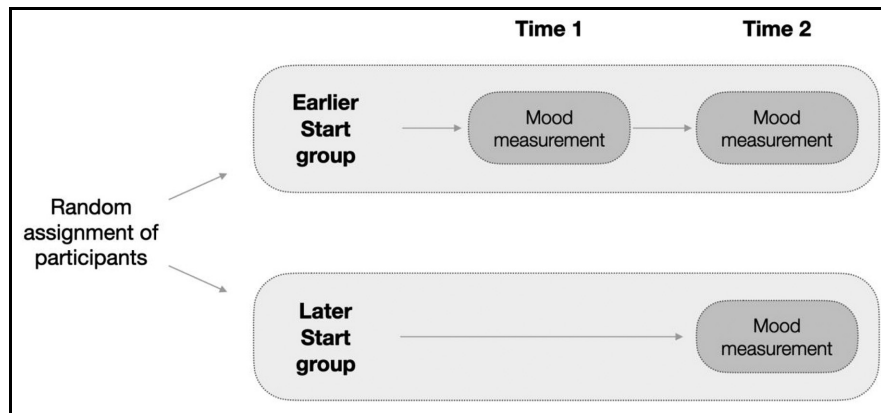Email: faridanvari.fa@gmail.com

Figure 1. Basic Experimental Design. Participants were recruited and randomly allocated to either the Earlier Start group or the Later Start group. Participants in the Earlier Start group completed the measures at both T1 and T2, and participants in the Later Start group completed the measures only at T2.

with people who had already responded earlier, especially for negative subjective experiences. The researchers called the phenomenon the initial elevation bias.

Shrout et al.'s (2018) results bring into question the validity of research findings that are based on self-report measures of affect, mood, and emotions, and of subjective experiences more generally. As such, the initial elevation phenomenon has serious implications for studies that are assessing the well-being of populations or evaluating the effectiveness of interventions and training programs. For example, when absolute levels are of interest, for instance in the epidemiology of mental health, the initial elevation phenomenon may lead to overestimated prevalence rates. Similarly, results from research assessing the subjective well-being of populations will be biased upward. Research using pre- and post-treatment measures will over- or underestimate the effects of treatment. Improvement in the control group, often attributed to placebo effects and regression to the mean, can also be a consequence of initial elevation. Research measuring longitudinal changes will also produce biased results. Indeed, the results of virtually all longitudinal studies may be affected to varying degrees. Moreover, if the initial elevation bias varies as a function of demographic characteristics, such as age, gender, or culture, then the relationships between these demographics and the affected measures will be biased.

However, more recently, Arslan et al. (2021) analyzed data from an experimental study with daily diary entries of more than 1,200 women recruited from German universities and various online platforms. They found no evidence of initial elevation. It is therefore unclear whether the initial elevation phenomenon is robust or whether it generalizes beyond Shrout et al.'s (2018) study and sample characteristics.

This situation poses a dilemma for researchers relying on self-reports. To mistakenly assume that there is no initial elevation might ignore a pervasive threat to the validity of many research results. On the contrary, assuming there is a bias when there isn't might cause researchers to waste valuable resources to mitigate the threat (e.g., Arslan et al., 2021; Shrout et al., 2018) or incorrectly reduce confidence in research findings. In the present paper, we report the results of three studies in which we tested (1) the robustness of the initial elevation phenomenon, (2) its most plausible mechanism, and (3) potential boundary conditions.

Methodological Paradigm of the Studies

For all three studies, we preregistered the hypotheses, methods, and analyses, and did not deviate from the preregistrations (Study 1: https://osf.io/xsr2q/?view_only = a5856 f24cac74820bc41428457a14747; Study 2: https://osf.io/v59 2b/?view_only = 09e88d46079c4baebd50ad62d4ef727e; Study 3: https://osf.io/6w43e/?view_only = 4bc1289b30ac49c6a52ec f4b7f636c50). For Study 2, we also preregistered additional exploratory, non-focal hypotheses which we report and discuss in the Supplemental Material. All raw data and analysis code in R, as well as the Supplemental Material with additional details on the samples and study materials, are on the Open Science Framework (https://osf.io/qtve3/?view_only = 6ad4c2bb65d843bfb1470b8959913630). We ran all studies using the Qualtrics survey software.

The basic design was a between-subjects experiment in which we recruited all participants on the same day and randomly allocated them to two groups: Earlier Start and Later Start. Participants in the Earlier Start group completed two surveys, one at Time 1 (T1) and one at Time 2 (T2). Participants in the Later Start group only completed the survey at T2 (see Figure 1). Because we recruited all participants on the same day and randomly allocated them to different start dates, there should be no difference in affective state between the Earlier Start and Later Start groups at T2. However, if there is an initial elevation on any measure, then participants in the Later Start group

should have higher ratings than participants in the Earlier Start group on the measure taken at T2.

Study 1

We designed Study 1 to examine whether the initial elevation phenomenon is robust. Whereas Shrout et al. (2018) used samples of U.S. college students and Arslan et al. (2021) used only German women, we recruited a diverse convenience sample from Prolific.co, an online platform often used for psychological research (Uittenhove et al., 2022). Ours is the largest sample, to date, to experimentally investigate the phenomenon. The preregistered confirmatory hypothesis, testing for whether the phenomenon is robust, was that there would be an initial elevation observed for in-the-moment reports of anxious mood state. We also preregistered three exploratory hypotheses examining whether there would be an initial elevation for self-reports of vigor, and arguably the most widely used measures of general positive and negative affect, the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988).

Methods

Procedure. Participants were recruited on a Monday, July 19, 2021. Participants read an information sheet, gave informed consent, and then were randomly allocated to either the Earlier Start or Later Start group (using Qualtrics randomization in "survey flow") and given instructions about when they would be contacted regarding participation. The day after recruitment (T1), we invited participants in the Earlier Start group to take the survey with the mood measures presented in the order listed in the Measures section below. One day after that (T2), we invited all participants (i.e., in both the Earlier Start and Later Start groups) to take the same survey. On the T2 survey, participants were additionally presented with two anchor items for an unrelated project. These anchor items were randomly varied to be either before or after the PANAS, asking participants to indicate how much more/less positive/negative they felt compared with when they completed the survey at T1. Given that they didn't do the T1 survey, participants in the Later Start group were asked to report how they felt compared with when they signed up to the recruitment study. Because of this, the Earlier and Later start groups' responses on the anchor items are not comparable and so these are not discussed further.

The invitations on both T1 and T2 were sent at around 12:00 BST and participants had until about 21:00 BST to complete the surveys. After each survey, participants were thanked for their time and partially debriefed (they were only *fully* debriefed after T2). Participants were paid £0.10 for the recruitment phase and £0.25 for each time they completed the 2-min survey. We used the Prolific.co pre-screening criteria to obtain participants' demographics.

Participants. We conducted statistical power analyses using G*power (Erdfelder et al., 1996) for two-tailed *t*-tests against zero, and the TOSTER package (Lakens et al., 2018) for the equivalence testing. We used a standardized effect size of $d = 0.16$ as the smallest effect size of interest, the lower median estimate of the initial elevation as reported by Shrout et al. (2018). For the two-tailed, independent samples *t*-tests to have 95% power to detect an effect size of Cohen's $d = 0.16$, with alpha $= .05$, we required a total sample size of 2,034. For the equivalence test to have 95% power to reject effect sizes of $-0.16 \geqslant d \geqslant 0.16$, with alpha $= .05$, we required a total sample size of 2,032. Anticipating attrition due to the longitudinal nature of the studies, we aimed to recruit a total of 2,300 participants. For each study, we ended up recruiting slightly over 2,300 participants because some people who signed up to the study timed-out such that Prolific didn't count these as part of the sample size even though their data were recorded in Qualtrics as having been assigned to a group. Given the increasing use of online participant samples, including from Prolific.co (Bohannon, 2016; Uittenhove et al., 2022), this provides an important investigation into how much the phenomenon affects studies that use these samples.

We recruited 2,305 participants. After attrition and following the preregistered exclusion criteria, there were a total of 1,856 participants for the analyses (Earlier Start $n = 903$; Later Start $n = 953$; gender details are in the Supplemental Material). Mean age of the sample was 26.76 years ($SD = 8.24$) ranging from 18 to 71 years. The attrition rate in the Earlier Start group (21.4%) was slightly higher than in the Later Start group (17.5%), $p = .021$.

Measures

Anxiety Scale. We measured anxious mood state using in-the-moment reports on a three-item scale (i.e., On edge, Uneasy, Anxious) validated in past research (Cranford et al., 2006) and found to be consistently affected by an initial elevation phenomenon (Shrout et al., 2018). Instructions for this and the other measures are in the Supplementary Materials. Participants responded to the three items, which were presented in random order, on 5-point rating scales, from $1 = $ *not at all* to $5 = $ *extremely* (the same 5-point scales were used also for the other three measures). The ratings for the three items were averaged for each participant. Cronbach's alpha for this scale was .83 at T1 and .85 at T2.

Vigor Scale. After the anxiety scale, participants gave in-the-moment reports on a three-item scale measuring vigor

Table 1. Results of Study 1.

Measure	M (SD) Late	M (SD) Early	Cohen's d [CI$_{95\%}$]	Inferential statistics
Anxiety	2.36 (1.04)	2.02 (1.00)	0.33 [0.24, 0.42]	$t_{(1853.7)}$ = 7.11, p < .001
Vigor	2.81 (0.93)	2.77 (0.98)	0.04 [−0.05, 0.13]	$t_{(1831.8)}$ = 0.83, p = .408
Positive affect	2.92 (0.83)	2.81 (0.91)	0.13 [0.03, 0.22]	$t_{(1816.5)}$ = 2.71, p = .007
Negative affect	2.01 (0.79)	1.81 (0.78)	0.26 [0.17, 0.35]	$t_{(1851.9)}$ = 5.60, p < .001

CI = confidence interval.
Note. The M (SD) Late and M (SD) Early columns present the means and standard deviations for the Later and Earlier start groups, respectively.

(i.e., Vigorous, Cheerful, Lively; Cranford et al., 2006; Shrout et al., 2018). Ratings for these three items were averaged to get a measure of vigorous mood state for each participant. Cronbach's alpha for this scale was .84 at T1 and .85 at T2.

PANAS. The PANAS is arguably the most widely used measure of positive and negative affect, with the original paper introducing and validating the measure currently having over 45,000 citations (Google Scholar, as of April 2022). Thus, we included the PANAS due to its scientific impact, which participants responded to after the vigor scale. The PANAS has 10 items measuring positive mood and 10 items measuring negative mood (the items are in the Supplemental Material). The positive and negative affect items were averaged for each participant to give a score for current positive and negative mood states, respectively. Cronbach's alpha for the positive affect subscale was .91 at both T1 and at T2. For the negative affect subscale, it was .89 at both T1 and at T2.

Exclusion/Inclusion Criteria. We preregistered four criteria for excluding/including participants from/in analyses: (1) participants had to be fluent in English, (2) only participants with complete data for all measures were included in analyses, (3) participants who gave the same rating for all items were excluded, and (4) only the first complete response for any one participant was included. Full details of the inclusion/exclusion criteria of this and the other studies are in the Supplemental Material.

Results

We used two-sided Welch's independent samples *t*-tests across all three studies. The descriptive and inferential statistics and standardized effect sizes for all measures are presented in Table 1 and Figure 2, and the distributions of the ratings are visualized in Figure 3.

The preregistered confirmatory test showed that there was an initial elevation on the anxiety scale. Participants in the Later Start group had statistically significantly higher
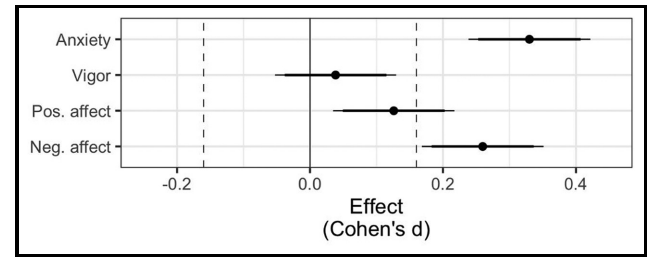
Figure 2. Effect Sizes in Study 1. The vertical dashed lines are the equivalence bounds, set at |*d* = 0.16|. The black dots represent the effect size point estimate for each measure and the thick and thin whiskers to either side of the black dots represent the 90% and 95% confidence intervals around the effect size, respectively. Pos. affect and Neg. affect are the positive and negative affect subscales of the PANAS, respectively.
Note. PANAS = Positive and Negative Affect Schedule.

ratings on the anxiety scale than participants in the Earlier Start group. We can therefore conclude that the initial elevation phenomenon is robust.

The preregistered exploratory test for the vigor scale showed that the initial elevation was smaller than the smallest effect size of interest. The difference in ratings on the vigor scale between participants in the Later and Earlier start groups was statistically nonsignificant. Moreover, the equivalence test was statistically significant, $t_{(1831.82)}$ = −2.62, p = .005, meaning that the effect on vigor fell within the equivalence bounds (|*d* = 0.16|).

The two other preregistered exploratory tests showed that there was an initial elevation for in-the-moment reports of positive and negative affect on the PANAS. Participants in the Later Start group had statistically significantly higher ratings on both the positive and negative affect subscales of the PANAS than participants in the Earlier Start group.

We conducted robustness checks to account for the difference in attrition rates between the groups using two approaches, including one very strict approach that input maximum anxiety ratings for 46 participants (to equalize the attrition rate between groups) in the Earlier Start group
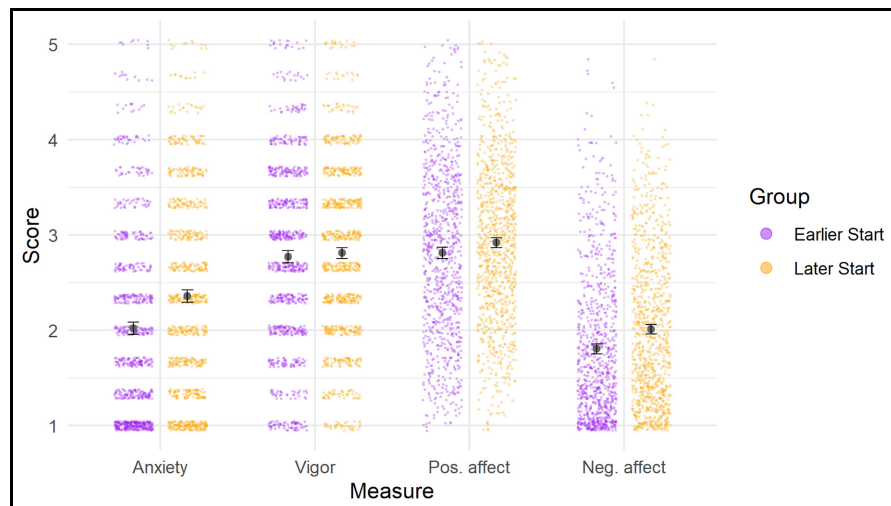
Figure 3. Distribution of Ratings in Study 1. Pos. affect and Neg. affect are the positive and negative affect subscales of the PANAS, respectively. Each coloured dot represents a participant's rating on that measure at T2. The dots are scattered around the possible values for each scale (e.g., anxiety had three items rated from 1-5 meaning that the possible values for any participant were 1, 1.33, 1.66, 2, and so on). The black dots and whiskers represent the means and 95% confidence interval around the mean, respectively. An initial elevation is present when the Later Start group has a higher mean than the Earlier Start group.
Note. PANAS = Positive and Negative Affect Schedule.

who responded at T1 but not at T2. The results of the robustness checks did not change the inferences (see Supplemental Material).

Discussion

The results of Study 1 make us confident that the initial elevation phenomenon is robust and not restricted to Shrout et al.'s (2018) study and design characteristics or to anxious mood. So what is the cause?

The most plausible mechanism is that there's a decrease in test-taking anxiety (Windle, 1955) or some change in feelings caused by taking part in the measurement process (French & Sutton, 2010; Knowles et al., 1996). Shrout et al. (2018, Study 3) argued that they ruled this mechanism out because they found an initial elevation for other-reports, when participants reported on their roommates' mental distress. But people's reports of others' feelings can also be driven, at least partly, by how they feel themselves (Trilla et al., 2021; see also "Social Projection" in Van Boven et al., 2013). Therefore, the initial elevation observed on other-reports may have been a reflection of the change in participants' own feelings.

Study 2

If the initial elevation phenomenon is driven purely by a change in feelings (i.e., a decrease in anxiety), then there should be a reverse effect on a scale that measures serenity, the opposite of anxiety. If people's feelings become *less* anxious over time, then they should also become *more* serene.

In addition, there should be an initial elevation for anxiety even if the anxiety scale is presented *only* at T2—that is, the Earlier Start group is also responding to the anxiety scale for the first time. Hence, if the initial elevation phenomenon is driven by a change in feelings, then at T2, compared with the Earlier Start group, the Later Start group should have *lower* ratings for serenity (H1a) and *higher* ratings for anxiety (H1b). In contrast, if the initial elevation phenomenon is not driven by a change in feelings *and* it affects positively valenced low arousal items, then we would observe an initial elevation phenomenon on the serenity scale (H2).

Furthermore, we included retrospective reports of positive and negative affect, expecting an initial elevation on these (H3a and H3b, respectively), and Big Five personality measures. Results of the preregistered tests on the Big Five were not statistically significant and are reported in the Supplemental Material.

Methods

Procedure. The procedure was exactly the same as Study 1 except that the anxiety scale was presented only in the T2 survey. Participants were recruited on a Monday, November 1, 2021, with T1 and T2 surveys being taken on Tuesday and Wednesday, respectively. Participants were paid £0.10 for the recruitment phase and £0.30 for each time they completed the 3-min survey. Participants responded to the measures in the order listed below (the measures reported in the Supplemental Material came after those listed here). We used the Prolific.co prescreening criteria to obtain participants' demographics and to prevent

Table 2. Results of Study 2.

Measure	M (SD) Late	M (SD) Early	Cohen's d [CI$_{95\%}$]	Inferential statistics
Serenity	3.33 (1.03)	3.37 (1.03)	−0.04 [−0.13, 0.05]	$t_{(1875.9)} = -0.83$, $p = .406$
Anxiety	2.30 (1.12)	2.36 (1.06)	−0.06 [−0.15, 0.04]	$t_{(1875.9)} = -1.20$, $p = .232$
Positive affect	4.32 (1.24)	4.32 (1.23)	0.001 [−0.09, 0.09]	$t_{(1876)} = -0.01$, $p = .991$
Negative affect	3.25 (1.03)	3.15 (1.07)	0.09 [−0.001, 0.18]	$t_{(1868.3)} = 1.95$, $p = .052$

CI = confidence interval.

Note. The *M (SD)* Late and *M (SD)* Early columns present the means and standard deviations for the Later and Earlier start groups, respectively.

participants who participated in Study 1 from taking part in Study 2.

Participants. We recruited 2,311 participants from Prolific.co. After attrition and following the preregistered exclusion criteria, there were a total of 1,879 participants for the analyses (Earlier Start *n* = 927; Later Start *n* = 952; gender details in the Supplemental Material). Mean age of the sample was 27.00 years (*SD* = 8.44) ranging from 18 to 84 years. Attrition rates between the groups were not significantly different (Earlier = 19.3% vs. Later = 17.4%, *p* = .252).

Measures. Full details of all measures, including instructions, are in the Supplemental Material.

Serenity Scale. At T1 and T2, participants first gave in-the-moment reports of serenity on the three-item subscale from the PANAS-X (i.e., At ease, Calm, Relaxed; Watson & Clark, 1994) using the same 5-point scales as in Study 1. These adjectives are also in the popular State-Trait-Anxiety-Inventory as anxiety-absent items (e.g., Marteau & Bekker, 1992). The ratings for the three items were averaged for each participant. Cronbach's alpha for this scale was .89 at T1 and .90 at T2.

Anxiety Scale. Only at T2, participants then gave in-the-moment reports on the same anxiety scale as in Study 1 (i.e., Uneasy, On edge, Anxious) on the same 5-point scales. Ratings for the three items were averaged for each participant. Cronbach's alpha was .88.

Retrospective Reports of Affect. At T1 and T2, participants gave retrospective reports of their feelings over the past month on items measuring positive and negative affect on 7-point frequency scales (1 = *never* to 7 = *always*). We used one item from each emotion category developed by Diener et al. (1995), so that we had two items for positive affect (Happiness, Affection) and four items for negative affect (Worry, Anger, Unhappiness, Shame). Ratings were averaged for the positive and negative affect scales separately. Positive affect had a Cronbach's alpha of .55 at T1 and .63 at T2. Cronbach's alpha for both positive affect and negative affect.

Exclusion/Inclusion Criteria. We preregistered the same exclusion/inclusion criteria as for Study 1 except that in Study 2 we also excluded participants who had participated in Study 1.

Results

The descriptive and inferential statistics and standardized effect sizes for all measures are presented in Table 2. H1a, H1b, and H2 were not supported. The difference in ratings at T2 between participants in the Later and Earlier start groups was not statistically significant for either the serenity or anxiety scales. Moreover, the difference between groups for both the serenity and anxiety scales fell within the equivalence bounds, as can be seen in Figure 4—the placement of the 90% confidence intervals relative to the equivalence bounds provides the same information as the equivalence tests such that if the interval falls within the equivalence bounds then the equivalence test is statistically significant. Therefore, the effects on the serenity and anxiety scales were not only statistically nonsignificant but also smaller than the smallest effect size of interest (i.e., |*d* = 0.16|). This suggests that the initial elevation phenomenon is unlikely to be driven purely by a change in feelings.

The inferential tests also did not support H3a or H3b, though the effect for retrospective reports of negative affect did not fall within the equivalence bounds.

Discussion

The results of Study 2 (H1a and H1b), when considered together with Shrout et al.'s (2018) findings for other-reports, suggest that the initial elevation phenomenon is not driven purely by a change in feelings. We can also be more confident that the phenomenon applies less strongly to items measuring pleasant, low arousal, mood states (H2). We found no evidence of initial elevation on retrospective reports of positive affect (H3a), and the initial elevation on retrospective reports of negative affect was not statistically significant (H3b). This was surprising because Shrout et al. (2018, Study 3) found that the phenomenon also applied to retrospective reports of mental distress. The Earlier Start group in our study may have remembered their ratings from T1, only one day ago, and subsequently gave consistent ratings at T2, which would attenuate initial elevation.
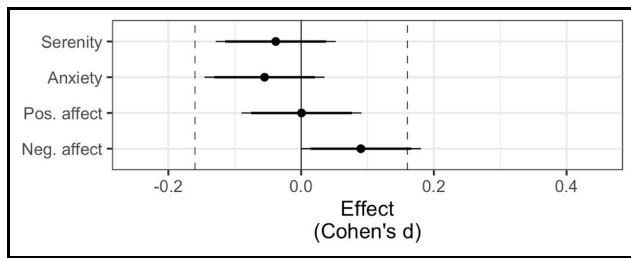
Figure 4. Effect Sizes in Study 2. The vertical dashed lines are the equivalence bounds, set at |*d* = 0.16|. The black dots represent the effect size point estimate for each measure and the thick and thin whiskers to either side of the black dots represent the 90% and 95% confidence intervals around the effect size, respectively. Pos. affect = positive affect; Neg. affect = negative affect; PoliticalSat = satisfaction with political leadership.

Note. Pos. affect = positive affect; Neg. affect = negative affect; PoliticalSat = satisfaction with political leadership.

Study 3

The question thus remains whether the phenomenon applies, beyond in-the-moment reports of affect, to retrospective reports. Knowing the boundary conditions is important because if the phenomenon also applies to retrospective reports then the scope of research affected is much broader.

Study 3 tested whether retrospective reports of negative mood are affected by initial elevation. We hypothesized that there would be initial elevation for self-reported symptoms of generalized anxiety disorder and depression (H1 and H2, respectively). We also preregistered non-focal analyses (that we report in the Supplemental Material). Finally, to test the robustness of Shrout et al.'s (2018) findings that extended the phenomenon to other subjective experiences, we hypothesized an initial elevation on a set of items asking participants to retrospectively report symptoms of physical illness (H3).

Methods

Procedure. The procedure was exactly the same as Study 1 except that (1) the T1 survey was given to participants in the Earlier Start group immediately after the recruitment procedure, and (2) the time between T1 and T2 was 2 weeks. Participants were recruited on a Tuesday, January 11, 2022, with the T1 survey being done immediately after recruitment. The T2 survey was done on Tuesday, January 25, exactly 2 weeks after the T1 survey. At T1, participants in the Later Start group were paid £0.10 and those in the Earlier Start group were paid £0.35 for doing the 3-min survey. At T2, all participants were paid £0.35. Participants responded to the three measures in the order listed below at both T1 and T2. After responding to the scales at T2, two additional items asked participants how often they felt anxious/depressed in the past 2 weeks compared with the 2

weeks before that. These items were included for an unrelated project but are used for exploratory analyses reported in the Supplemental Material. We used the Prolific.co prescreening criteria to obtain participants' demographics. For this study, we used a different author's Prolific account and so we did not use the prescreening to prevent participants from Studies 1 and 2 from participating. We found that 48 participants who participated in Study 1 and 78 who participated in Study 2 also participated in Study 3. Removing these participants did not change the inferential results and the effect sizes stayed largely the same (except for slight increases such that the effect size for each measure was $d = 0.13$). In the results section, we report the results without excluding these participants.

Participants. We recruited 2,322 participants from Prolific.co. After attrition and following the preregistered exclusion criteria, there were a total of 1,550 participants for the analyses (Earlier Start $n = 788$; Later Start $n = 762$; gender details in the Supplemental Material). Mean age of the sample was 26.62 years ($SD = 7.45$) ranging from 18 to 73 years. Attrition rates between the groups were not significantly different (Earlier $= 30\%$ vs. Later $= 32.8\%$, $p = .153$).

Measures. In this study, we included two scales that are widely used to assess symptoms of generalized anxiety disorder and depression. According to Google Scholar, as of February 2022, the seven-item Generalized Anxiety Disorder (GAD-7) validation paper (Spitzer et al., 2006; see also Löwe et al., 2008) has over 14,000 citations, and the eight-item Patient Health Questionnaire (PHQ-8) validation papers (Kroenke et al., 2009; Kroenke & Spitzer, 2002) have over 7,000 citations combined. The PHQ-8 has one (suicidality) item removed from the PHQ-9. The validation paper of the PHQ-9 (Kroenke et al., 2001) has over 26,000 citations. A third scale asked participants how much they experienced eight physical symptoms. For each of the three scales, participants rated how much they experienced each symptom over the last 2 weeks, from 0 = *Not at all* to 4 = *Nearly every day*. For the items and full instructions, see Supplemental Material. Ratings on the GAD-7 and PHQ-8 items were summed for each participant to get a score representing the severity of anxiety and depression, respectively. Ratings for physical illness symptoms were averaged for each participant. The GAD-7 had Cronbach's alpha of .86 at T1 and .87 at T2. For the PHQ-8, it was .85 at both T1 and at T2. For the physical symptoms scale, it was .67 at T1 and .69 at T2.

Exclusion/Inclusion Criteria. We preregistered the same exclusion/inclusion criteria as for Study 1 except that in Study 3

Table 3. Results of Study 3.

Measure	M (SD) Late	M (SD) Early	Cohen's d [CI$_{95\%}$]	Inferential statistics
GAD-7	8.63 (5.01)	8.02 (4.74)	0.12 [0.03, 0.22]	$t_{(1535.6)} = 2.46, p = .014$
PHQ-8	9.17 (5.69)	8.47 (5.35)	0.13 [0.03, 0.23]	$t_{(1534.3)} = 2.49, p = .013$
Physical symptoms	0.67 (0.48)	0.62 (0.46)	0.12 [0.02, 0.22]	$t_{(1539.9)} = 2.32, p = .020$

CI = confidence interval; GAD = Generalized Anxiety Disorder; PHQ = Patient Health Questionnaire.
Note. The M (SD) Late and M (SD) Early columns present the means and standard deviations for the Later and Earlier start groups, respectively.

we also included a criterion limiting the number of studies participants should have previously submitted on Prolific (see Supplemental Material for details).

Results

Supporting H1 to H3, we found a statistically significant initial elevation on each of the three scales (see Table 3). At T2, the Later Start group reported more symptoms of anxiety, depression, and physical illness, as compared with the Earlier Start group. Figure 5 shows that the 90% confidence intervals did not fall within the equivalence bounds. Based on these results, we can conclude that there was an initial elevation phenomenon on the three retrospective reports in this study.

Discussion

There was an initial elevation on retrospectively reported symptoms of generalized anxiety disorder (H1), depression (H2), and physical illness (H3). In light of Shrout et al.'s (2018) results showing (also a smaller) initial elevation on retrospective reports of mental distress and physical symptoms, the current findings make us more confident that the phenomenon applies to retrospective reports of (1) negative mood and (2) negative subjective experiences other than affect, mood, and emotions.

General Discussion

Although a recent high-profile publication suggests that an initial elevation phenomenon on self-reports of subjective experiences biases the results of many studies (Shrout et al., 2018), more recent publications report little evidence for such a phenomenon (Arslan et al., 2021; Cerino et al., 2022). However, Arslan et al. (2021) had many methodological differences from Shrout et al. that could explain the results (e.g., the first response of each participant to the items being compared occurred several days into the diary study for some participants), and the results from Cerino et al. are from within-person analyses (i.e., no experimental controls) that could be confounded with (and hence masked by) the effects of time. In the largest experiments to date, designed specifically for testing the initial elevation phenomenon, we found that self-reports of negative (and
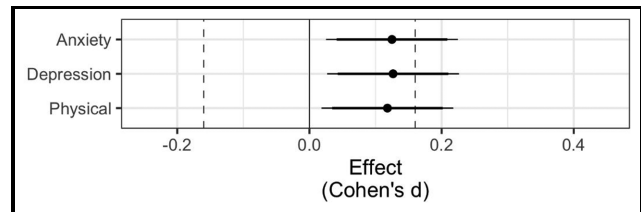

Figure 5. Effect Sizes in Study 3. The vertical dashed lines are the equivalence bounds, set at |d = 0.16|. The black dot represents the effect size point estimate for each measure and the thick and thin whiskers to either side of the black dots represent the 90% and 95% confidence intervals around the effect size, respectively.

some positive) subjective experiences are indeed affected by an initial elevation. Consistent with Shrout et al., we found that the phenomenon occurs more strongly for negative subjective reports than for positive subjective reports, though Shrout et al. reported larger median effect sizes for negative subjective reports than we found here (see Supplemental Material under heading "Significance of Effect Sizes in Main Analyses" for interpretation of effect sizes in present studies in reference to other studies and also for other exploratory analyses.)

Our results indicate that the initial elevation phenomenon is robust and generalizes to online samples that are increasingly used in social science research (Bohannon, 2016; Uittenhove et al., 2022). These findings converge with findings from other studies using U.S. college students (Shrout et al., 2018). Nonetheless, our results may not generalize to other populations. For example, because we used an online participant pool by Prolific.co, many participants had substantial survey taking experience (Study 1 did not have data on number of surveys each participant had completed; Study 2: $M = 49$, $SD = 82$, median = 23, range = 1–1,749; Study 3: $M = 44$, $SD = 45$, median = 23, range = 1–209). On the contrary, out of various online participant pools, Prolific.co reports lower levels of careless responding than competitors (Peer et al., 2022). To further reduce careless responding, we excluded participants who gave the same response to every item. In addition, a benefit of online participant pools is that they are demographically more diverse than student research pools (Berinsky et al., 2012; Gosling et al., 2004). Future research should

investigate the initial elevation phenomenon in samples from other populations and examine whether the size of the phenomenon varies as a function of survey experience.

Moreover, although we used a variety of validated and widely used scales, the generalizability of our findings should be further examined by looking at a greater variety of scales and scale types. For example, research should investigate the initial elevation phenomenon on scales with fewer and more scale points, with different labels on the scale points, with more items measuring the subjective experience, and validated measures of different subjective experiences.

We found that the phenomenon applies to in-the-moment reports of mood states, such as anxiety and positive and negative affect as measured by the PANAS (Study 1), and to retrospective reports of mental and physical health symptoms (Study 3). Study 2 ruled out that the effect is driven purely by a change in feelings. The initial elevation phenomenon thus poses a threat to the validity of many research designs that use self-reports of negative subjective experiences.

Measuring people's subjective experiences with rating scales is increasingly common in social science research. As such, the initial elevation phenomenon threatens the validity of a broad range of behavioral, social, and cognitive research areas. Our results should spur wider recognition of this phenomenon—a type of measurement reactivity (French et al., 2021)—as well as efforts to understand it and curtail its negative effect on research findings.

Author Contributions

F.A. conceptualized the studies and initiated all aspects of the work. E.E., J.O., and I.K.S. provided feedback in the conceptualization process for the studies and on the drafts of the paper. R.C.A. and M.E. provided feedback in the conceptualization process for Studies 2 and 3 and provided feedback on the final draft of the paper. J.O. and R.C.A. checked the main analyses and conducted additional analyses of the data. All authors read and approved the final version of the text.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

Ethics

The research complies with all necessary ethical requirements. Ethical approval was granted by the Faculty of Human Sciences Ethics Commission at the University of Cologne. Informed consent was obtained from all participants prior to participation.

ORCID iDs

Farid Anvari (iD) https://orcid.org/0000-0002-5806-5654
Iris K. Schneider (iD) https://orcid.org/0000-0003-0915-0809

Supplemental Material

Supplemental material for this article is available online.

References

Arslan, R. C., Reitz, A. K., Driebe, J. C., Gerlach, T. M., & Penke, L. (2021). Routinely randomize potential sources of measurement reactivity to estimate and adjust for biases in subjective reports. *Psychological Methods*, *26*(2), 175–185. https://doi.org/10.1037/met0000294

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Bohannon, J. (2016). Mechanical Turk upends social sciences. *Science*, *352*(6291), 1263–1264. https://doi.org/10.1126/science.352.6291.1263

Brantley, P. J., Cocke, T. B., Jones, G. N., & Goreczny, A. J. (1988). The Daily Stress Inventory: Validity and effect of repeated administration. *Journal of Psychopathology and Behavioral Assessment*, *10*(1), 75–81. https://doi.org/10.1007/BF00962987

Cerino, E. S., Schneider, S., Stone, A. A., Sliwinski, M. J., Mogle, J., & Smyth, J. M. (2022). Little evidence for consistent initial elevation bias in self-reported momentary affect: A coordinated analysis of ecological momentary assessment studies. *Psychological Assessment*, *34*(5), 467–482. https://doi.org/10.1037/pas0001108

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

Dukes, D., Abrams, K., Adolphs, R., Ahmed, M. E., Beatty, A., Berridge, K. C., Broomhall, S., Brosch, T., Campos, J. J., Clay, Z., Clément, F., Cunningham, W. A., Damasio, A., Damasio, H., D'Arms, J., Davidson, J. W., de Gelder, B., Deonna, J., de Sousa, R., & . . . Sander, D. (2021). The rise of affectivism. *Nature Human Behaviour*, *5*(7), 816–820. https://doi.org/10.1038/s41562-021-01130-8

Diener, E., Smith, H., & Fujita, F. (1995). The personality structure of affect. *Journal of Personality and Social Psychology*, *69*(1), 130–141. https://doi.org/10.1037/0022-3514.69.1.130

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 1–11. https://doi.org/10.3758/BF03203630

French, D. P., Miles, L. M., Elbourne, D., Farmer, A., Gulliford, M., Locock, L., Sutton, S., & McCambridge, J., & The MERIT Collaborative Group. (2021). Reducing bias in trials from reactions to measurement: The MERIT study including developmental work and expert workshop. *Health Technology Assessment*, *25*(55), 1–72. https://doi.org/10.3310/hta25550

French, D. P., & Sutton, S. (2010). Reactivity of measurement in health psychology: How much of a problem is it? What can be done about it? *British Journal of Health Psychology*, *15*(3), 453–468. https://doi.org/10.1348/135910710X492341

Ganzach, Y., & Bulmash, B. (2021). The effect of serial day on the measurement of positivity and emotional complexity in diary studies. *European Journal of Social Psychology*, *51*(7), 1213–1225. https://doi.org/10.1002/ejsp.2809

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*(2), 93–104. https://doi.org/10.1037/0003-066X.59.2.93

Iachina, M., & Bilenberg, N. (2012). Measuring reliable change of emotional and behavioural problems in children. *Psychiatry Research*, *200*(2–3), 867–871. https://doi.org/10.1016/j.psychres.2012.06.023

Knowles, E. S., Coker, M. C., Scott, R. A., Cook, D. A., & Neville, J. W. (1996). Measurement-induced improvement in anxiety: Mean shifts with repeated assessment. *Journal of Personality and Social Psychology*, *71*(2), 352–363. https://doi.org/10.1037/0022-3514.71.2.352

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, *32*(9), 509–515. https://doi.org/10.3928/0048-5713-20020901-06

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1–3), 163–173. https://doi.org/10.1016/j.jad.2008.06.026

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., & Herzberg, P. Y. (2008). Validation and standardization of the Generalized Anxiety Disorder screener (GAD-7) in the general population. *Medical Care*, *46*(3), 266–274. https://doi.org/10.1097/MLR.0b013e318160d093

Lucas, C. P., Fisher, P., Piacentini, J., Zhang, H., Jensen, P. S., Shaffer, D., Dulcan, M., Schwab-Stone, M., Regier, D., & Canino, G. (1999). Features of interview questions associated with attenuation of symptom reports. *Journal of Abnormal Child Psychology*, *27*(6), 429–437. https://doi.org/10.1023/A:1021975824957

Marteau, T. M., & Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State—Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology*, *31*(3), 301–306. https://doi.org/10.1111/j.2044-8260.1992.tb00997.x

Milich, R., Roberts, M. A., Loney, J., & Caputo, J. (1980). Differentiating practice effects and statistical regression on the Conners Hyperkinesis Index. *Journal of Abnormal Child Psychology*, *8*(4), 549–552. https://doi.org/10.1007/BF00916506

Neprash, J. A. (1936). The reliability of questions in the Thurstone personality schedule. *The Journal of Social Psychology*, *7*(2), 239–244. https://doi.org/10.1080/00224545.1936.9921665

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Piacentini, J., Roper, M., Jensen, P., Lucas, C., Fisher, P., Bird, H., Bourdon, K., Schwab-Stone, M., Rubio-Stipec, M., Davies, M., & Dulcan, M. (1999). Informant-based determinants of symptom attenuation in structured child psychiatric interviews. *Journal of Abnormal Child Psychology*, *27*(6), 417–428. https://doi.org/10.1023/A:1021923808118

Reynolds, B. M., Robles, T. F., & Repetti, R. L. (2016). Measurement reactivity and fatigue effects in daily diary research with families. *Developmental Psychology*, *52*(3), 442–456. https://doi.org/10.1037/dev0000081

Ribera, J. C., Canino, G., Rubio-Stipec, M., Bravo, M., Bauermeister, J. J., Alegría, M., Woodbury, M., Huertas, S., Guevara, L. M., Bird, H. R., Freeman, D., & Shrout, P. E. (1996). The Diagnostic Interview Schedule for Children (DISC-2.1) in Spanish: Reliability in a Hispanic population. *Journal of Child Psychology and Psychiatry*, *37*(2), 195–204. https://doi.org/10.1111/j.1469-7610.1996.tb01391.x

Sharpe, J. P., & Gilbert, D. G. (1998). Effects of repeated administration of the Beck Depression Inventory and other measures of negative mood states. *Personality and Individual Differences*, *24*(4), 457–463. https://doi.org/10.1016/S0191-8869(97)00193-1

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092. https://doi.org/10.1001/archinte.166.10.1092

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

Trilla, I., Weigand, A., & Dziobek, I. (2021). Affective states influence emotion perception: Evidence for emotional egocentricity. *Psychological Research*, *85*(3), 1005–1015. https://doi.org/10.1007/s00426-020-01314-3

Uittenhove, K., Jeanneret, S., & Vergauwe, E. (2022). *From lab-based to web-based behavioural research: Who you test is more important than how you test* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/uy4kb

Van Boven, L., Loewenstein, G., Dunning, D., & Nordgren, L. F. (2013). Changing places. In J. M. Olson, & M. P. Zanna

(Eds.), *Advances in experimental Social psychology* (*Vol. 48*, pp. 117–171). Elsevier. https://doi.org/10.1016/B978-0-12-40 7188-9.00003-X

Watson, D., & Clark, L. A. (1994). *The PANAS-X: Manual for the positive and negative affect schedule—Expanded form* [Data set]. University of Iowa. https://doi.org/10.17077/48vt-m4t2

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

Windle, C. (1954). Test-Retest Effect on Personality Questionnaires. *Educational and Psychological Measurement*, *14*(4), 617–633. https://doi.org/10.1177/001316445401400404

Windle, C. (1955). Further Studies of Test-Retest Effect on Personality Questionnaires. *Educational and Psychological Measurement*, *15*(3), 246–253. https://doi.org/10.1177/0013164455 01500304

Author Biographies

Farid Anvari is interested in the measurement of psychological constructs and the underlying assumptions of such measurements. His research has focused on self-reported affect, mood, emotions, and well-being, and the biases involved in these measures.

Emir Efendić is an assistant professor at the School of Business and Economics in Maastricht University in the Netherlands. His research focuses on judgment and decision-making.

Ruben Arslan studies genetic and hormonal causes of why people differ and change, especially their personality, intelligence, and sexuality. Open, reproducible science is important to him, so he focuses on methods, measures, and transparency and dabbles in research software development.

Jerome Olsen is a psychologist who is interested in things that have to do with data, R, open science, and official statistics. He is now the head of department at the Federal Statistical Office (Destatis) in Bonn, Germany.

Malte is a behavioral psychologist and assistant professor of Psychology of Human Technology Interaction at Ruhr University Bochum. He studies human learning in various contexts, such as the contingencies of behaviors in academic research (meta science), and human interaction with & effects of technology.

Iris Schneider is a professor of Social Psychology at the TU Dresden. She studies ambivalence and mixed feelings in judgment and (social) decision-making. More on her work can be found at www.irisschneider.nl.