Abstract

Research shows that people's self-reports may be biased by an initial elevation phenomenon in which ratings are higher the first time that people take a survey as compared to the second and subsequent times. Apart from the fact that this phenomenon exists, and that it might bias ratings for negative subjective experiences more strongly than positive ones, little else is known. In the present study, we examined whether the initial elevation phenomenon occurs for commonly used trait measures, such as ratings on personality inventories and life satisfaction. We hypothesized that the initial elevation phenomenon may be associated with the (un)desirability of the content of the self-report items such that scores for undesirable facets would show initial elevation and scores for desirable facets would show the reverse. We tested this in an online convenience sample (N = 3,329) using 5 facets of a personality inventory and a single item measure of life satisfaction. Our hypotheses were not supported. Our findings suggest that at least for online convenience samples, ratings on personality inventories and life satisfaction are not strongly impacted by initial elevation.

Introduction

Previous research shows that self-reports may be biased by an initial elevation phenomenon. Controlled experiments in which participants are, for example, randomly allocated to respond to survey items either on 2 measurement occasions (T1 and T2) or only once (T2) have found that people's ratings tend to be higher, on average, for those who are taking the survey for the first time as compared to a second time (Anvari et al., 2023; Shrout et al., 2018). Shrout et al. (2018) and Anvari et al. (2023) found that the initial elevation phenomenon occurs for various self-reports, including momentary ratings of anxiety, general negative affect, and general positive affect, and retrospective self-reports of mental and physical health symptoms. Although the phenomenon seems robust for measures of these constructs, many questions remain. One question with practical relevance for researchers is whether the initial elevation phenomenon occurs more generally, for measures of trait-like constructs. For example, it remains unclear whether the initial elevation phenomenon occurs for self-reports on personality inventories and for life satisfaction, both of which are widely used in the social sciences, and the mechanisms driving the phenomenon are unknown.

Previously proposed mechanisms for the initial elevation phenomenon involved some process in which people's feelings would change in the time between completing the first and subsequent surveys. Perhaps the most obvious explanation for any changes observed between two measurement occasions are the confounds of time, whereby some event causes changes in the experience being measured. But there have also been other mechanisms proposed. For example, the initial elevation phenomenon has been explained as a reduction in test-taking anxiety (e.g., Windle, 1955), a general change in the feelings of participants caused by taking part in the measurement process (French & Sutton, 2010; Knowles et al., 1996), or as a result of sampling bias whereby people high in anxiety are selected into a study for the first survey and the most anxious people subsequently dropout of the study or become less anxious by the time the second survey is administered (e.g., Arslan et al., 2021; Iachina & Bilenberg, 2012; Milich et al., 1980). However, these studies did not include experiments with control groups to test the proposed mechanisms.

Recent studies with experimental controls have ruled out these previously proposed explanations for the phenomenon. For example, Shrout et al. (2018) randomly allocated people to different groups with the only difference being when the groups would start

participating in the longitudinal study (e.g., one group would start the daily diary study on Day 1 and another group would start the study on Day 22). They could therefore compare the different groups on their survey responses to the same questions on the same day (e.g., Day 22), but whereas one group would be completing the survey for the first time on that day, the other group would have completed the survey at least once before. Shrout et al. found that people who were taking the survey for the first time (e.g., the group starting on Day 22) had consistently higher ratings, particularly for negative subjective experiences such as anxiety, than people who had started the study earlier (e.g., the group who started on Day 1). Moreover, Shrout et al. found that the initial elevation occurred even when people were reporting the psychological distress of their roommates, so that the phenomenon could not be explained by the confounds of time or any other mechanism causing a change in the participants' feelings.

Further evidence against the previously proposed mechanisms came from a study conducted by Anvari et al. (2023). These researchers recruited a large sample of participants from an online participant pool on Day 0 and randomly allocated them to two groups. One group completed the survey on Days 1 and 2, whereas the other group completed the survey only on Day 2. An initial elevation would be present when on Day 2, the group responding to the survey for the first time would have higher ratings than the group responding to the survey for the second time. In Study 1, this is exactly what Anvari et al. observed: Participants completing the survey for the first time on Day 2 reported higher levels of anxiety than those who were completing the survey for the second time. In Study 2, Anvari et al. made a slight change whereby the survey on Days 1 and 2 consisted of a 3-item measure of serenity, presumably the opposite of anxiety, and only the Day 2 survey included the 3-item measure of anxiety that they used in Study 1. Anvari et al. reasoned that if the initial elevation phenomenon was caused by a change in feelings, such as a reduction in anxiety, then on Day 2 there should be an initial elevation on the anxiety scale even though both groups were responding to this scale for the first time, and a corresponding reverse effect on the serenity scale. There were no statistically significant differences, with effects for both scales falling within the equivalence bounds ($d = |0.16|$). Taken together with the findings in Shrout et al. (2018), the initial elevation phenomenon is unlikely to be caused by real changes in feelings.

In trying to identify alternative, plausible mechanisms for the initial elevation phenomenon we examined the data patterns reported in previous work. Anvari et al., (2023, Supplemental) reported exploratory analyses showing that there was an initial elevation phenomenon for personality items that could be classified as being undesirable, but not for items that could be classified as desirable. We analyzed these data on an item-by-item basis and found that some undesirable items showed an initial elevation phenomenon and that some desirable items showed a reversed effect. We hypothesized that the initial elevation phenomenon may therefore be related to the (un)desirability of the personality traits.

Present Research

For the study we report here, we hypothesized that personality facets we identified as being undesirable would show an initial elevation phenomenon. In contrast, we expected that facets that are desirable would show a reverse initial elevation. For those facets closer to neutral, we predicted an effect falling within the equivalence bounds. Details of how we selected the personality facets are in the methods section. Finally, we also included a single item measure of life satisfaction to test whether this widely used measure is impacted by the initial elevation phenomenon. We recruited a large sample of participants and randomly allocated them to two groups. One group completed the measures directly after recruitment (T1) as

well as in 2 weeks' time (T2; this is "the T1-T2 group"). The other group completed the measures only at T2 (this is "the T2-only group").

We were interested in examining the role of the (un)desirability of the content of the personality measures in the initial elevation phenomenon; namely, that people may at first report higher levels of undesirable content and lower levels of desirable content. Therefore, our hypotheses for the personality measures were not concerned with whether items were reverse scored. Rather, our hypotheses had to do with whether higher or lower scores on the personality measures indicated more or less of an (un)desirable thing. With this in mind, we considered that an initial elevation phenomenon would be evident on a personality measure if the T2-only group had higher scores than the T1-T2 group at T2, after scoring the measures in accordance with the scoring key described in the methods section. And we expected that with the personality measures so scored, those which were high in undesirability would show an initial elevation phenomenon, those high in desirability would show a reverse effect, and those which were low on (un)desirability would have an effect falling within the equivalence bounds. We included life satisfaction for the purposes of another, unrelated study, and so we this measure was included only incidentally which we, nonetheless, preregistered a hypothesis for, expecting an initial elevation. We therefore had the following preregistered hypotheses for responses at T2:

Hypothesis 1a: The T2-only group who are responding for the first time to the survey items will have *higher* scores for the highly undesirable Self-Consciousness facet, compared to the T1-T2 group who are responding for the second time (i.e., there would be initial elevation);

Hypothesis 1b and 1c: The T2-only group will have *lower* scores than the T1-T2 group for the highly desirable Assertiveness and Self-Efficacy facets, respectively (i.e., there would be a reverse effect, with later elevation);

Hypothesis 1d and 1e: The difference in scores between the T2-only group and the T1-T2 group will fall within the equivalence bounds of Cohen's $d = 0.16$ for the relatively neutral Sympathy and Artistic Interests facets, respectively;

Hypothesis 2: The T2-only group will have *lower* scores than the T1-T2 group for life satisfaction (i.e., there would be a reverse effect, with later elevation).

Methods

The hypotheses, methods, and analyses were all preregistered on the OSF: https://osf.io/y29ch/?view_only=78a1c4743c844c1a9fe6305f8b7e3f4f. The data, analysis code, and materials (i.e., Qualtrics printouts) are also on the OSF: https://osf.io/e2yj9/?view_only=25286d900b0448a3adb19f321b41403d.

Participants and Procedure

Our smallest effect size of interest was $d = 0.16$, the same as in Anvari et al. (2023) and the lower median estimate found in Shrout et al. (2018). For each independent samples, two-sided, Welch's *t*-test to have 95% power to detect an effect size of $d = 0.16$, we required 1,445 participants for each condition, i.e. 2,890 participants in total. To reject effects at least as large as $d = |0.16|$ with equivalence tests at 95% power, we required 2,984 participants. Given the longitudinal nature of the study, we expected attrition. Therefore, we recruited 4,000 participants from Prolific.co on 5th July 2023 (Wednesday). After removal of duplicate entries, we ended up with a sample of 3,995 participants recruited. We randomly allocated participants to the T1-T2 group or the T2-only group. The T1-T2 group completed the survey items immediately after giving informed consent (T1) and then were told that they would be

contacted to do another survey in 2 weeks. The T2-only group were told that they would be contacted in 2 weeks to do the survey (i.e., they completed none of the survey items at T1). In 2 weeks (T2), on the same day of the week as the T1 survey (i.e., Wednesday 19th July), both groups were invited to do the survey. The T2 survey was kept open for about 24 hours. At T1, participants in the T2-only group were paid £0.10 and participants in the T1-T2 group were paid £0.40. At T2, all participants were paid £0.40. After removal of duplicate entries, we had a sample of 3,329 participants who completed the survey at T2.

Using Prolific's records, rather than collecting demographic data in our own study, 1,693 participants reported sex as female, 1,614 as male, 7 preferred not to say, and 9 participants for whom the sex-data on Prolific had expired. Participants in the final sample had a mean age of 36.67 years (*SD* = 12.75 years), ranging from 18 to 123 years (given that Jeanne Calment, the oldest verified person to have lived, died at 122 years old, the 123 year old person in our data is likely an entry error; the next oldest person was 81). the T1-T2 group who were responding for the second time at T2 consisted of 1,662 participants, and the T2-only group of 1,667.

Measures

Personality

To inform our selection of undesirable, desirable, and neutral personality facets, we started with the results of analyses of the facets in the NEO-PI-R reported by Schimmack (2019). Table 3 in Schimmack (2019) shows each facet's loading onto an evaluative bias factor. The evaluative bias factor reflects whether a facet is undesirable (negative loading), desirable (positive loading), or more neutral (low or zero loading). The NEO-PI-R is not freely available. Therefore, we used the free 120-item IPIP (Maples et al., 2014), which has items, facets, and domains that correspond well with the NEO-PI-R.

To make sure that the facets from the 120-item IPIP loaded onto the evaluative bias factor similarly to the corresponding facets of the NEO-PI-R reported by Schimmack (2019), we conducted factor analyses on data from Bainbridge et al. (2022) who used the same 120-item IPIP that we used. Bainbridge et al. had two studies. Their Study 2 (2a and 2b) had the larger sample size. We used the data from this study to tailor the measurement model, by adding cross-loadings and allowing facet residuals to correlate, and identify a good fitting model. We extracted from this model the loading of each facet onto the evaluative bias factor. We then validated the model using the data from Bainbridge et al.'s Study 1 and, again, extracted each facet's loading onto the evaluative bias factor. The R code and data for these analyses are on the OSF link provided above which also includes an excel file containing each facet and its factor loading onto the evaluative bias factor in our models using Bainbridge et al's data. We selected the facets from the IPIP which had loadings onto the evaluative bias factor similar to the loadings of the corresponding facets in Schimmack (2019).

We thus selected the undesirable Self-Consciousness facet (N4); the desirable Assertiveness (E3) and Self-Efficacy facets (C1; in Schimmack's 2019 inventory the latter is called Competence); and the relatively neutral Sympathy (A6) and Artistic Interests facets (O2; in Schimmack's 2019 inventory these were called Tender-Mindedness and Aesthetics, respectively). The items from each of the selected facets are presented below. The evaluative bias factor loadings for the facets we selected in the validation model were: Self-Consciousness (-.55); Assertiveness (.65); Self-Efficacy (.61); Sympathy (.21); and Artistic Interests (.09). We were therefore confident that the facets we selected from the 120-item IPIP had appropriate loadings on the evaluative bias factor, so that two facets were highly desirable, one was highly undesirable, and two were neutral or low on (un)desirability.

The personality items were used in both the T1 and T2 surveys. Participants were asked, "How much do you agree or disagree with each of the following statements?:

N4: Self-Consciousness
I find it difficult to approach others.
I am easily intimidated.
I am not embarrassed easily.*
I am able to stand up for myself.*

E3: Assertiveness
I take charge.
I try to lead others.
I take control of things.
I wait for others to lead the way.*

O2: Artistic Interests
I see beauty in things that others might not notice.
I do not like art.*
I do not like poetry.*
I do not enjoy going to art museums.*

C1: Self-Efficacy
I complete tasks successfully
I excel in what I do.
I handle tasks smoothly.
I know how to get things done.

A6: Sympathy
I sympathize with the homeless.
I feel sympathy for those who are worse off than myself.
I suffer from others' sorrows.
I am not interested in other people's problems.*

The above personality items were presented in random order on a single page and rated: 1 = disagree strongly, 2 = disagree a little, 3 = neither disagree nor agree, 4 = agree a little, 5 = agree strongly. Ratings on the items for each facet were averaged for each participant, after reverse scoring the relevant items. (* Indicates the item should be reverse scored.)

Life Satisfaction

At both T1 and T2, after the personality items and on a separate page, we included a single item measure of life satisfaction as used in the German Socio-Economic Panel (Goebel et al., 2019):

"How satisfied are you with your life, all things considered?", rated from 0 = completely dissatisfied, to 10 = completely satisfied.

Additional Measures

At T2 only, on the same page as the life satisfaction item, we asked participants 3 further questions for another study unrelated to this one:

"Remember 2 weeks ago when you first signed up to this study. How satisfied were you with your life at that time, all things considered?", rated from 0 = completely dissatisfied, to 10 = completely satisfied.

"Compared to 2 weeks ago, would you say that you are less satisfied with your life now, more satisfied now, or about the same?", 1 = much less satisfied, 2 = a little less satisfied, 3 = about the same, 4 = a little more satisfied, 5 = much more satisfied.

"What rating did you give 2 weeks ago for how satisfied you were with your life? (If you didn't do the survey 2 weeks ago, choose "n/a".)", rated from 0 = completely dissatisfied, to 10 = completely satisfied, and n/a being the 11th response option.

We do not report these last 3 items any further in this paper.

Attrition Analyses

We assessed potential differences in attrition between the two groups. Overall attrition, going from T1 to T2, was 16.6% (665 participants completed the T1 recruitment phase/survey but did not complete the T2 survey). Of these, 335 were the T1-T2 group and 330 were from the T2-only group. The attrition rate for the T1-T2 group was 16.8% and for the T2-only group it was 16.5%, and this difference was not statistically significant, X^2 = 0.03, p = .8595. Therefore, attrition is unlikely to explain any potential differences between the T2-only and the T1-T2 groups.

We also assessed potential differences in the T1 responses on the measures for the T1-T2 group between those who completed the T2 survey and those who did not complete the T2 survey. Although this would not inform us about potential differences on the measures between the T2-only and the T1-T2 groups, it could shed light on potential differences for the within-group analyses reported in the "Exploratory Analyses" section of the results. For example, if people who dropped out had scores that consistently reflected poorer adjustment (e.g., higher Self-Consciousness, lower Life Satisfaction and Self-Efficacy, and so on), then such a systematic pattern could be used to infer why there might or might not be differences in scores between T1 and T2. There were only two measures (out of six) on which T1 scores were statistically significantly different between those who dropped out and those who did not. Those who dropped out (M = 3.93, SD = 0.67) had lower scores on Self-Efficacy than those who did not (M = 4.01, SD = 0.67), $t_{(477.12)} = 2.00$, p = .0462. In contrast, those who dropped out (M = 6.56, SD = 0.1.87) had higher scores on life satisfaction than those who did not (M = 6.26, SD = 1.97), $t_{(495.62)} = -2.66$, p = .0080. The differences on the other measures were not statistically significant (all ps > .10) and the two significant differences may be chance findings given the number of multiple tests (the Bonferonni corrected alpha for six comparisons is .008). There was therefore not a clear pattern that could be used to shed light on the results of the within-group analyses described further below.

Results

Preregistered Analyses

For all analyses, ratings given at T2 were the dependent variables. With 6 tests (Hypotheses 1a-1e, and 2), the corrected alpha was preregistered to be (.05/6) .008. We tested Hypotheses 1a-1e and 2 using independent samples, two-sided, Welch's *t*-tests. The *t*-tests for the hypotheses that were not significant were followed up with equivalence tests. Hypotheses 1d and 1e were to be tested using equivalence tests to examine whether the effect size falls within the equivalence bounds of d = -0.16 and 0.16. The means, standard deviations, and *t*-test results comparing the groups at T2 are presented in Table 1.

Table 1.

Measure	*M* (*SD*) T2-only	*M* (*SD*) T1-T2	Cohen's *d* [$CI_{95\%}$]	Inferential Statistics

Self-Consciousness	2.72 (0.86)	2.68 (0.89)	0.04 [-0.03, 0.11]	$t_{(3321.6)} = 1.10$, $p = .2733$
Assertiveness	3.24 (0.89)	3.29 (0.93)	-0.05 [-0.12, 0.01]	$t_{(3322.2)} = 1.55$, $p = .1216$
Self-Efficacy	3.95 (0.67)	4.00 (0.66)	-0.07 [-0.14, -0.01]	$t_{(3326.2)} = 2.14$, $p = .0321$
Sympathy	3.73 (0.75)	3.78 (0.76)	-0.07 [-0.14, 0.000]	$t_{(3325.9)} = 1.96$, $p = .0504$
Artistic Interests	3.70 (0.93)	3.76 (0.93)	-0.07 [-0.13, 0.002]	$t_{(3326.9)} = 1.91$, $p = .0560$
Life Satisfaction	6.22 (2.00)	6.23 (2.05)	-0.01 [-0.08, 0.06]	$t_{(3324.5)} = 0.25$, $p = .8031$

Note. The *M* (*SD*) columns present the means and standard deviations for each group. Cohen's *d*s were calculated using the cohen.d() function from the "effsize" package (version 0.8.1; Torchiano, 2020) in R, with pooled standard deviation and Hedge's correction. A positive Cohen's *d* indicates that the T2-only group had higher mean scores than the T1-T2 group (i.e., initial elevation) and a negative Cohen's *d* indicates the reverse.

Hypothesis 1a was that the T2-only group would have higher mean ratings for Self-Consciousness than the T1-T2 group. This hypothesis was not supported. The equivalence test showed that the difference between the groups fell within the equivalence bounds, $t_{(3321.6)} = 3.52$, $p = .0002$.

Hypothesis 1b and 1c were that the T2-only group would have lower mean ratings than the T1-T2 group for Assertiveness and Self-Efficacy, respectively. Neither of these were supported. The equivalence tests showed that the effects fell within the equivalence bounds for Assertiveness, $t_{(3322.25)} = 3.07$, $p = .0011$, as well as for Self-Efficacy, $t_{(3326.2)} = 2.47$, $p = .0067$.

Hypothesis 1d and 1e were that the difference between the T2-only group and the T1-T2 group would fall within the equivalence bounds of $d = 0.16$ for ratings of Sympathy and Artistic Interests, respectively. The equivalence test was statistically significant for both Sympathy, , $t_{(3325.89)} = 2.66$, $p = .0040$, and Artistic Interests, $t_{(3326.92)} = 2.70$, $p = .0034$.

Finally, Hypothesis 2 was that the T2-only group would have lower mean ratings than the T1-T2 group for life satisfaction. This hypothesis was also not supported, and the equivalence test showed that the difference fell within the equivalence bounds, $t_{(3324.49)} = 4.37$, $p < .0001$.

Exploratory Analyses

In exploratory, non-preregistered analyses, we also examined whether the T1-T2 group (n = 1,662) showed any change in ratings going from T1 to T2. With 6 measures and therefore 6 tests, we used the corrected alpha of .008 for these exploratory analyses. An initial elevation would be evident if the scores at T1 are higher than the scores at T2. The means, standard deviations, and one-sample t-test results comparing T1 with T2 for the T1-T2 group are presented in Table 2.

Table 2.

Measure	***M* (*SD*) T1**	***M* (*SD*) T2**	**Cohen's *d* [$CI_{95\%}$]**	**Inferential Statistics**
Self-Consciousness	2.68 (0.89)	2.68 (0.89)	0.003 [-0.03, 0.03]	$t_{(1661)} = 0.19$, $p = .8457$
Assertiveness	3.37 (0.92)	3.29 (0.93)	0.08 [0.05, 0.11]	$t_{(1661)} = 5.58$, $p < .0001$

Self-Efficacy	4.01 (0.67)	4.00 (0.66)	0.01 [-0.02, 0.04]	$t_{(1661)} = 0.48$, $p = .6297$
Sympathy	3.79 (0.76)	3.78 (0.76)	0.01 [-0.02, 0.04]	$t_{(1661)} = 0.89$, $p = .3752$
Artistic Interests	3.77 (0.91)	3.76 (0.93)	0.01 [-0.02, 0.04]	$t_{(1661)} = 0.82$, $p = .4111$
Life Satisfaction	6.26 (1.97)	6.23 (2.05)	0.01 [-0.01, 0.04]	$t_{(1661)} = 1.03$, $p = .3025$

Note. The *M* (*SD*) columns present means and standard deviations at T1 and T2 for the T1-T2 group. Cohen's *d*s were calculated using the cohen.d() function from the "effsize" package (version 0.8.1; Torchiano, 2020) in R, with pooled standard deviation and Hedge's correction. A positive Cohen's *d* indicates that the T1 scores were higher than the T2 scores (i.e., initial elevation) and a negative Cohen's *d* indicates the reverse.

The only within-group comparison that was significant was for Assertiveness, for which there was an initial elevation. However, this was in the direction opposite to what would be expected if our proposed mechanism of (un)desirability of the content of the measure was the explanation, since Assertiveness was highly desirable and which we therefore expected to show a reverse effect (see Hypothesis 1b). All of the other within-group differences were centred almost directly over zero.

Discussion

We tested whether the initial elevation phenomenon occurs for self-report ratings on several facets of a personality inventory and for life satisfaction, and whether the phenomenon corresponds with the (un)desirability of the facets' content. The pattern of results we hypothesized was not supported. There was a statistically nonsignificant small initial elevation on the undesirable facet of self-consciousness, and nonsignificant small reverse effects on the 2 desirable facets and the 2 relatively neutral facets. For life satisfaction, the effect was centred almost exactly on zero. All effects fell within the equivalence bounds. Therefore, if there is an initial elevation phenomenon (or reversal) that occurs on the facets of the personality inventory and for life satisfaction in participants from convenience samples, it is likely to be smaller than $d = |0.16|$. Our smallest effect size of interest was based on results from past research examining the initial elevation phenomenon using similar methodological designs. Shrout et al. (2018), reported $d = 0.16$ as the smallest median effect size from their studies, and Anvari et al. (2023) used that as their smallest effect size of interest. Researchers conducting future studies may select an effect size relevant for them, based on their own justifications.

In exploratory analyses we examined whether there was support for our hypotheses at the within-person level, by comparing the ratings the T1-T2 group gave at T1 against T2. Five of the effects centred almost exactly on zero, with the Assertiveness facet showing an initial elevation; which was in the opposite direction to the predicted effect based on the desirability of the facet. Because these within-person comparisons were exploratory and less causally informative because of the lack of randomization, we do not draw any conclusions about the one statistically significant effect.

The simplest interpretation of our results is that personality inventories and life satisfaction ratings, such as those in the present study, are not affected by or less prone to an initial elevation phenomenon as compared to the affect and symptom measures used in past research (e.g., Anvari et al., 2023; Shrout et al., 2018). And, furthermore, this parsimonious

interpretation of our results suggests that the initial elevation phenomenon is unrelated to the (un)desirability of the content of the measures.

There could be various explanations as to why personality inventories and life satisfaction ratings may be less prone to an initial elevation phenomenon. The ratings for both the personality inventories and the life satisfaction ratings are given on bipolar scales. For the personality items, participants reported how much they disagreed-agreed with each statement, and for life satisfaction participants reported how much they were dissatisfied-satisfied. Experimentally controlled studies showing evidence for the initial elevation phenomenon have used unipolar scales, such as where participants give ratings from 1 = not at all, to 5 = extremely (Anvari et al., 2023; Shrout et al,. 2018). One could therefore speculate that bipolar scales may, for some unknown reason, be immune to initial elevation.

However, we urge caution against drawing any strong conclusions about personality inventories and life satisfaction ratings being immune to initial elevation across all contexts. Participants in our sample had been approved, on Prolific, for completing a mean number of 555 (SD = 546) studies; median of 368. It is therefore plausible that we may not have observed initial elevation on the personality and life satisfaction measures due to participants having had experience with completing similar measures previously. If previous survey experience immunized participants against the initial elevation phenomenon, we would have to additionally assume that: (i) personality inventories and life satisfaction scales are included in online surveys more often than affect measures, since initial elevation has, in past research, been found for samples from Prolific using measures of momentary anxiety and general affect, as well as retrospective reports of mental and physical health symptoms (Anvari et al., 2023); and (ii) a substantial part of our sample had indeed participated in a study that included personality and life satisfaction measures. Some experimentally controlled studies have reported that the initial elevation phenomenon may last at least up to 2 months, with effect sizes declining over time. People who completed the measures 2 months ago tend to still have lower ratings than people completing the measures for the first time now (Shrout et al., 2018; see also Windle, 1954).

There is also research on panel conditioning effects (or instrumentation, as discussed by Baird et al., 2010) that suggests initial elevation may last substantially longer than 2 months. For example, Wooden and Li (2013) examined 10 waves of a yearly panel survey of Australian households, testing for changes in life satisfaction and mental health. For life satisfaction, they found an initial elevation in people's ratings on the first wave as compared to the second wave, even after including other potential influencing variables in their model, noting that the initial elevation effect was small (0.06 points, on an 11-point scale). Their analyses of the mental health measure showed no initial elevation when their model included the influencing variables, except when they examined only the women in their data. They observed that women's scores increased in successive waves, suggesting a reverse initial elevation; but higher scores on their measure were indicative of better mental health so that in the first wave women reported poorer mental health. This is partly consistent with the findings of experimentally controlled studies in which mental health measures show initial elevation, such that people reported elevated mental and physical health symptoms in the first survey as compared to the second (Anvari et al., 2023; Shrout et al., 2019). In what may perhaps be considered a better test of initial elevation on panel research, Van Landeghem (2012) used the German SOEP and Swiss Household Panel survey to compare life satisfaction ratings given by respondents from refreshment samples being added to the panels against the ratings given by existing panel members completing the surveys on a yearly basis. Van Landeghem found that those in the refreshment samples completing the survey for the first time had higher life satisfaction ratings than those who had completed the surveys

already (differences ranged from 0.16 to 0.64, on an 11-point scale), even after controlling for age differences between the refreshment and the existing samples.

Moreover, experimentally controlled studies have also been used to examine panel conditioning effects. For example, Torche et al. (2012) conducted a panel study of adolescents, randomly allocating students to two groups. Both groups completed the survey at both T1 and T2 which were separated by a year. One group's survey included questions about substance use at both T1 and T2, whereas the other group's survey included the substance use questions only at T2. Torche et al. found that the group responding to the substance use questions for the first time at T2 had a higher rate of reporting substance use than those responding for the second time (i.e., there was an initial elevation effect; but see Halpern-Manners et al., 2014, where an effect in the opposite direction was found for other illicit behaviors among adults). Similarly, in another panel study that was conducted every 2 months, participants were randomly allocated to complete additional weekly surveys or to continue with just the bi-monthly surveys; although there was no effect on most measures, those who completed surveys less frequently had higher ratings for feeling depressed (i.e., an initial elevation; Cornesse et al., 2023). Taken together, panel conditioning effects in the longitudinal survey literature show that initial elevation may be observed even when surveys are separated by one year.

This means that, because our participants were experienced survey-takers who had likely completed personality inventories and life satisfaction scales relatively recently, we cannot rule out the role of the (un)desirability of the content of the measures in the initial elevation phenomenon, nor can we conclude that personality inventories and life satisfaction ratings are generally immune in samples of participants with no previous experience with these measures. Nonetheless, our findings still have practical relevance. An increasing number of studies in psychological science use opt-in online samples, such as those supplied by Prolific (Bohannon, 2016; Uittenhove et al., 2022). It is therefore important for researchers using such samples to have some indication of whether their measures might be impacted by initial elevation.

In attempting to explain the difference between our results in the present study and past research using online convenience samples finding an initial elevation, one could also argue that our sample may have contained a small but not insignificant proportion of bogus respondents. Kennedy et al. (2020) found that bogus respondents in opt-in online samples (4% to 7%), like the one we used, tend to say "Yes" to most questions regardless of content, perhaps to increase their chances of being screened in for a subsequent study (see also Mercer & Lau, 2023). It is conceivable that professional survey takers have learned that feigning negative affect and psychopathology, but not undesirable personality, increases their odds of being screened into subsequent studies. However, several of the samples used by Shrout et al. (2018) were student samples, which should be immune to such concerns, and still showed the initial elevation phenomenon.

In conclusion, our results show that at least for samples from Prolific.com, and possibly other convenience samples commonly used by psychology researchers (e.g., MTurk and Cloud Research), ratings on personality inventories and life satisfaction are not very strongly impacted by the initial elevation phenomenon. Although these measures may be impacted by initial elevation in samples of participants who do not routinely participate in surveys, our findings show that researchers using convenience samples may not need to worry too much about the initial elevation phenomenon on personality inventories and life satisfaction scales.

References

Anvari, F., Efendić, E., Olsen, J., Arslan, R. C., Elson, M., & Schneider, I. K. (2023). Bias in Self-Reports: An Initial Elevation Phenomenon. Social Psychological and Personality Science, 14(6), 727–737. https://doi.org/10.1177/19485506221129160

Bainbridge, T. F., Ludeke, S. G., & Smillie, L. D. (2022). Evaluating the Big Five as an organizing framework for commonly used psychological trait scales. Journal of Personality and Social Psychology, 122(4), 749–777. https://doi.org/10.1037/pspp0000395

Baird, B.M., Lucas, R.E. & Donnellan, M.B. Life Satisfaction Across the Lifespan: Findings from Two Nationally Representative Panel Studies. Soc Indic Res 99, 183–203 (2010). https://doi.org/10.1007/s11205-010-9584-9

Cornesse, C., Blom, A., Sohnius, M. L., González Ocanto, M., Rettig, T., & Ungefucht, M. (2023). Experimental Evidence on Panel Conditioning Effects when Increasing the Surveying Frequency in a Probability-Based Online Panel. *Survey Research Methods*, 17(3), 323–339. https://doi.org/10.18148/srm/2023.v17i3.7990

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schroder, C., & Schupp, J. (2019). The German Socio-Economic Panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik*, *239*(2), 345–360. Retrieved from http://hdl.handle.net/10419/200973

Halpern-Manners, A., Warren, J. B., Torche, F. (2014). Panel Conditioning in a Longitudinal Study of Illicit Behaviors, *Public Opinion Quarterly*, 78(3), 565–590. https://doi.org/10.1093/poq/nfu029

Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J., & Asare-Marfo, D. (2020). Assessing the risks to online polls from bogus respondents. *Pew Research Center*, *18*. Retrieved from: https://www.pewresearch.org/methods/wp-content/uploads/sites/10/2020/02/PM_02.18.20_dataquality_FULL.REPORT.pdf

Maples, J. L., Guan, L., Carter, N. T., & Miller, J. D. (2014). A test of the International Personality Item Pool representation of the revised NEO Personality Inventory and development of a 120-item IPIP-based measure of the Five-Factor Model. *Psychological Assessment*, *26*, 1070-1084. https://doi.org/10.1037/pas0000004

Mercer, A., & Lau, A. (2023). Comparing Two Types of Online Survey Samples. *Pew Research Center*. Retrieved from: https://www.pewresearch.org/methods/2023/09/07/comparing-two-types-of-online-survey-samples/

Schimmack, U. (2019, August 14). What lurks beneath the Big Five?. Replicability-Index. https://replicationindex.com/category/evaluative-bias/

Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clave´ l, 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

Torche, F., Warren, J. R., Halpern-Manners, A., Valenzuela, E. (2012). Panel conditioning in a longitudinal study of adolescents' substance use: Evidence from an experiment. *Social Forces*, 90(3), 891–918. https://doi.org/10.1093/sf/sor006

Torchiano M (2020). effsize: Efficient Effect Size Computation. R package version 0.8.1, <https://CRAN.R-project.org/package=effsize>. https://doi.org/10.5281/zenodo.1480624.

Van Landeghem, B. (2012). Panel conditioning and self-reported satisfaction: Evidence from international panel data and repeated cross-sections.

Van Landeghem, B. (August 2012). Panel conditioning and self-reported satisfaction: Evidence from international panel data and repeated cross-sections. *SOEPpaper No. 484*, http://dx.doi.org/10.2139/ssrn.2146594

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

Wooden, M., Li, N. Panel Conditioning and Subjective Well-being. Soc Indic Res 117, 235–255 (2014). https://doi.org/10.1007/s11205-013-0348-1