Routinely randomize potential sources of measurement reactivity to estimate and adjust for biases in subjective reports

Arslan, Ruben C.; Reitz, Anne K.; Driebe, Julie C.; Gerlach, Tanja M.; Penke, Lars

Published in:
Psychological Methods

DOI:
10.31234/osf.io/va8bx
10.1037/met0000294

Publication date:
2021

Document Version
Early version, also known as pre-print

Routinely Randomize Potential Sources of Measurement Reactivity to Estimate and Adjust for Biases in Subjective Reports

Arslan, R. C.[1], Reitz, A. K.[2], Driebe, J. C.[3], Gerlach, T. M.[3,4], & Penke, L.[3,4]

1 Center for Adaptive Rationality, Max Planck Institute for Human Development
2 Department of Developmental Psychology, Tilburg University
3 Biological Personality Psychology, Georg Elias Müller Institute of Psychology, University of Göttingen
4 Leibniz ScienceCampus Primate Cognition

Abstract: With the advent of online and app-based studies, researchers in psychology are making increasing use of repeated subjective reports. The new methods open up opportunities to study behavior in the field and to map causal processes, but they also pose new challenges. Recent work has added initial elevation bias to the list of common pitfalls; here, higher negative states (i.e., thoughts and feelings) are reported on the first day of assessment than on later days. This article showcases a new approach to addressing this and other measurement reactivity biases. Specifically, we employed a planned missingness design in a daily diary study of more than 1,300 individuals who were assessed over a period of up to 70 days to estimate and adjust for measurement reactivity biases. We found that day of first item presentation, item order, and item number were associated with only negligible bias: items were not answered differently depending on when and where they were shown. Initial elevation bias may thus be more limited than has previously been reported or it may act only at the level of the survey, not at the item level. We encourage researchers to make design choices that will allow them to routinely assess measurement reactivity biases in their studies. Specifically, we advocate the routine randomization of item display and order, as well as of the timing and frequency of measurement. Randomized planned missingness makes it possible to empirically gauge how fatigue, familiarity, and learning interact to bias responses.

Keywords: measurement reactivity, experience sampling, repeated measures, planned missingness

Author Note: An earlier version of this manuscript was shared as a preprint on PsyArXiv (doi:10.31234/osf.io/va8bx). We thank the participants in our diary study and the Leibniz ScienceCampus Göttingen, which partially funded the study. Thanks go to Jojanneke Bastiaansen, Joanne Chung, Fabian Dablander, Jonas Haslbeck, Rich Lucas, and John Zelenski for interesting discussions on measurement reactivity at the SIPS 2019 conference. We thank Emorie Beck and Aidan Wright for helpful reviews. We thank Susannah Goss for scientific editing. All remaining errors are ours. Correspondence concerning this article should be addressed to Ruben C. Arslan, Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: ruben.arslan@gmail.com

Repeated subjective reports, in which individuals provide multiple reports on their thoughts, feelings, and/or behaviors within a short time span, are becoming increasingly popular across nearly all branches of psychological science, including clinical, social, personality, and health psychology. They have a wide range of applications (e.g., in research on individual development, well-being, and health) and are often used to examine causal processes, developmental change, and individual variability therein. A major reason for the growing popularity of repeated subjective reports is that this type of data has become relatively easy to assess now that digital devices such as smartphones and wearables are ubiquitous. Repeated subjective reports offer a window onto everyday life and causal processes, and they open up new possibilities for tailoring interventions to individuals' unique needs (Bolger & Laurenceau, 2013).

Because interest in specific substantive questions often outstrips research on measurement issues, many measurement and design decisions are made ad hoc and not evaluated rigorously. As research has become increasingly reliant on repeated subjective reports, various potential validity issues have been raised. In addition to random measurement error, systematic response biases have been discussed as a challenge to validity (Knowles & Condon, 1999). In particular, researchers have found that the reported severity of negative states, such as anxiety and depression, decreases across repeated reports (Knowles et al., 1996; Sharpe & Gilbert, 1998). An "attenuation effect," in which later responses are biased downward because of measurement reactivity, where the act of measurement affected people's later responses, and a selection bias in which people are more likely to enroll in a study if, for example, they are particularly anxious, and then regress to the mean were both seen as potential causes (Knowles et al., 1996; Shrout et al., 2017).

Recently, however, Shrout et al. (Shrout et al., 2017) showed that reports of subjective negative states tend to be biased upward when first assessed, not biased downward in later assessments. Participants in diary and experience sampling studies were randomized to different starting dates. The authors showed that participants' initial reports on subjective negative states were biased upward (initial elevation), and ruled out selection bias and regression to the mean as an explanation because the start date was randomised. If, as suggested by Shrout et al., the discrepancies are due to an initial elevation bias, this is bad news for research using subjective reports: This kind of bias would affect even non-repeated self-reports, perhaps the most frequent method of data collection in social science. In clinical psychology, it could potentially explain unexpected "improvements" in symptoms in control groups, exacerbating the better-studied biases of regression to the mean and placebo response. The bias would also violate the assumption of stationarity of means over the study period, which is common in most time series models (e.g., it could lead to unwarranted conclusion of cross-lagged effects that would disappear upon detrending). It would also mean that cross-sectional studies of negative subjective states would always overestimate the mean relative to repeated panel studies of the same population. Given these wide-ranging consequences, Shrout et al. called for further investigation of the initial elevation bias.

Shrout et al. (2017) randomized the starting dates of the entire repeated surveys. Therefore, we do not yet know the bias occurs at the item level, the survey level, or

both—that is, whether people give elevated responses when they see a new item for the first time or when they begin a new survey. If the bias operates at item level, it would also affect new questions in panel studies and would be even harder to overcome. Shrout et al. found that items about negative but not positive states were affected, suggesting that not all items are affected equally. But because only study starting dates were randomized in their studies, when the study was started was confounded with when an item was first seen.

On the basis of their findings, Shrout et al. proposed two countermeasures against initial elevation bias: (a) dropping initial observations and (b) familiarizing subjects with the survey before starting the assessment. These strategies are expensive and dropping observations would incur the loss of partially valid information. For new studies, we urge caution regarding these proposed countermeasures and instead recommend the design decision of randomizing potential causes of bias when planning a study (as done by Shrout et al. themselves). Dropping observations may be a passable robustness check when working with existing data, if there is a reason to expect initial elevation bias. However, randomization allows researchers to estimate and adjust for bias resulting from reactivity.[1] This is not only less wasteful, but also compatible with another piece of widely ignored best practice advice, namely, to implement planned missingness designs (Condon, 2018; Graham et al., 2006; Revelle et al., 2016; Silvia et al., 2014).

Many researchers who collect repeated subjective reports are keen to reduce waste and redundancy in their studies. They face the dilemma of wanting to keep their surveys brief in order to avoid drop-outs and to reduce fatigue, while not wanting to compromise on the breadth and number of constructs assessed. We posit that both this efficiency problem and the problem of potential measurement-related biases caused by item order, initial elevation, question familiarity, workload, and other forms of measurement reactivity can be addressed by one design decision made before data acquisition—namely, not showing all items on all days, but instead randomly selecting a subset of items to be shown in random order each day. This approach would allow surveys to be kept brief, expected biases to be estimated, and the necessary statistical adjustments for non-negligible biases to be made. Missing values that result from this approach are ignorable (Rubin, 1976); their handling usually requires no greater statistical expertise than researchers already need for the analysis of multilevel data. Changing questions and their order on a daily basis can also help to keep monotony at bay, preventing participants from responding “on autopilot” and reducing drop-out (Silvia et al., 2014). A planned missingness design for repeated subjective reports may thus reduce systematic missingness by reducing both participant fatigue and the contingency between responding to later items and fatigue (i.e., the fact that participants are more fatigued and less motivated when responding to later items; (Palen et

[1] Familiarizing subjects with questionnaires and other measurement instruments is presumably worthwhile for many samples and procedures. Doing so can unmask ambiguous or poorly phrased items or help uncover potential misunderstandings in less survey-experienced samples. For some instruments (e.g., some structured clinical interviews), explaining items and allowing questions to the interviewer is explicitly part of the procedure. Still, especially when such familiarization procedures are costly to implement, they are also perfect candidates for randomization. After all, individual, non-automatable effort in data collection often trades off against sample size in a very direct way.

al., 2008).

In this study, which re-used data from a diary study (Arslan et al., 2016) employing a planned missingness design, we used a method to estimate initial elevation bias that made it possible to examine the effects of when an item was first seen independent of when the study was started. We make the case that routinely estimating and adjusting for biases related to measurement reactivity, as demonstrated here, is cheap, easy, and desirable.

Methods

The diary study was designed to investigate changes in subjective states across the menstrual cycle. It is also ideally suited to investigate initial elevation bias and other biases related to measurement reactivity, because whether and in which order most items were shown was randomized in a simple planned missingness design (Silvia et al., 2014).

Recruitment and Incentives

We recruited participants between June 2016 and January 2017 through various online channels (i.e., the online platform psytests.de, advertisements on okCupid.com and Facebook, mass mailing lists of German-speaking university students) and by directly inviting suitable candidates who had taken part in previous studies with similar recruitment strategies. When recruitment stagnated, the study was additionally presented in a first-year psychology lecture. Data collection ended in May 2017. The incentives for taking part in the study were direct payment (of an amount ranging from €25 to €45),[2] the chance of winning prizes with a total value of €2,000[3] or, for University of Goettingen students, course credit. For all three incentives, the magnitude of the reward depended on the regularity of participation. As a further incentive, every participant received personalized graphical feedback at the end of the study.

Study Structure

Women participated in an online study named “Alltag und Sexualität” [Daily Life and Sexuality], which was implemented and automated using the survey framework formr.org (Arslan et al., 2019). The study was introduced as an online diary aiming to examine the interaction of sexuality, psychological well-being, experience of romantic relationships, and everyday experiences. The study had six main stages; here, we focus on the repeated diary measures. After completing consent forms, participants filled out a demographic questionnaire, the responses to which were used to decide whether they would be paid or entered in a lottery. Once they had been informed how they would be rewarded, participants completed a personality questionnaire, which was irrelevant for the current study as it was a single assessment. A day later, the online diary study began. Over a period of 70 days, women received an email invitation at 5 pm (they additionally received text message reminders if they had disclosed their mobile phone number and missed

[2] Only women fulfilling certain sample criteria were offered direct payment. These were being under the age of 50, being heterosexual, having a regular menstruation and being pre-menopausal, and not having taken any hormonal or psychoactive medication or hormonal contraception in the last 3 months. Additionally, women were paid only if they were not trying to get pregnant and had not been pregnant and/or breastfeeding within the last 3 months.

[3] The lottery prizes included an iPhone, an iPad, and forty €20 Amazon vouchers.

several diary entries). The online diary could be filled out until 3 am the following morning and included questions about their mood, daily activities, and sexuality. Items were randomized within grouped blocks of varying sizes. The items central to the main questions of the study were shown every day; items of lesser importance appeared randomly 20–80% of the time. Subsequent to the diary study, participants completed a social network questionnaire and a final follow-up questionnaire that assessed whether important changes in relationship and contraception status had occurred during the diary period (both of which are irrelevant to this analysis).

Data Used in the Present Analysis

We used 60,982 daily diary entries (mainly closed-ended questions) reported by 1,345 women over up to 70 days (M = 45 days, SD = 22). The participating women were between 18 and 61 years of age (M = 26.5, SD = 7.2) and had on average 15.1 years of education (SD = 4.8). Two thirds (67%) were students, 30% were employed. Nine percent were married, two percent engaged, 49% were in a committed relationship, and 32% were single, with the remainder being in non-committed relationships. Eleven percent had children. Subjects gave their informed consent (survey studies are exempt from ethics committee approval under German regulations). The present analysis focuses on responses to six items about loneliness, irritability, risk taking, self-esteem, stress, and mood, which were presented on the first page of the online diary (see Table 1). Each day, a random subset of these items was shown on the first page, in random order.[4] We randomized item selection and order on a daily basis and separately for each participant. Items were first shuffled, then pseudorandom numbers were drawn from a uniform distribution to determine which items would be shown (see Table 2 for group sizes). This procedure was implemented in formr.org using R (Arslan et al., 2019). Participants responded to each item on a 5-point Likert scale ranging from “less than usual” to “more than usual.” Pole labels were placed to the left and right of five equally-sized blank buttons.

The following variables were randomized: the number of times an item had previously been seen (conditional on adjusting for day number in the diary), item order, number of items shown on that day, and identity of the preceding item(s).

[4] For the subset of women who were in a relationship, there was one more randomised item "I was happy with my relationship." on the first page. Although we do not report this item as an outcome here to keep things simple, it is of course part of the number of items displayed, the item order, and the preceding item identity.

Table 1. Items Investigated in the Present Analysis

Item	N[women]	N[days]	Days/ Woman	Mean	SD	% shown
I felt lonely.	1305	24384	19	1.40	1.14	40
I was easily irritated.	1295	24538	19	1.61	1.12	40
I was prepared to take risks.	1265	12165	10	1.80	0.95	20
I was satisfied with myself.	1334	48923	37	2.11	0.97	80
I was stressed out.	1310	24435	19	1.82	1.18	40
My mood was good.	1334	48921	37	2.19	1.03	80

Note. Items were presented in German. The sample sizes differ across items due to randomization. *N*[women] = number of women who saw the item at least once. *N*[days] = number of days the item was shown. Days/Woman = number of days a woman saw the item on average. % shown = nominal and empirical percentage of the 60,982 days in the diary study on which the item was shown. See also Supplementary Figure 2 for histograms.

Table 2. The group sizes by randomized first day

	First day of item presentation							
Item	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7+**
I felt lonely.	575 (44%)	278 (21%)	160 (12%)	110 (8%)	66 (5%)	38 (3%)	22 (2%)	56 (4%)
I was easily irritated.	513 (40%)	281 (22%)	168 (13%)	107 (8%)	79 (6%)	48 (4%)	33 (3%)	69 (5%)
I was prepared to take risks.	251 (20%)	176 (14%)	165 (13%)	109 (9%)	101 (8%)	69 (5%)	68 (5%)	329 (26%)
I was satisfied with myself.	1078 (81%)	190 (14%)	43 (3%)	14 (1%)	4 (0%)	3 (0%)	1 (0%)	4 (0%)
I was stressed out.	530 (40%)	275 (21%)	159 (12%)	112 (9%)	73 (6%)	63 (5%)	39 (3%)	62 (5%)
My mood was good.	1078 (81%)	177 (13%)	52 (4%)	19 (1%)	3 (0%)	6 (0%)	0 (0%)	3 (0%)

Note. Each cell above shows how many women saw which item for the first time on which day. For the items with higher probability of being shown on each day, group sizes level off quickly. Diary days were counted beginning with zero.

We selected the general mood and state items presented in Table 1 for this analysis because they are typical of the items used in psychological research and include positive as well as negative subjective states. As a robustness check, we also examined some of the other diary items that were more specific to the research questions of the original study. Specifically, we selected five items that were not about subjective states, but tapped how participants spent their time; these items were again answered on a response scale from "less than usual" to "more than usual." To vary the response scales, we also chose six items on partnered sexual desire that were answered on a 5-point response scale from "inaccurate" to "very accurate" and three items on partner jealousy that were answered on a 5-point response scale from "not at all" to "very much."

Analysis

Our analyses and the code used to produce all figures and tables are fully documented on OSF (https://osf.io/7y6ag/). The data used cannot be fully anonymized and will be shared with scientists who agree not to attempt re-identification. We stored encrypted, access-controlled datasets on OSF and uploaded checksums to an immutable, and time-stamped component (https://osf.io/z2afv). Analyses were performed in R 3.5.2 (R Core Team, 2018) with the help of the tidyverse (Wickham, 2014), and lme4 (Bates et al., 2015) packages. Reproducible reports were generated using knitr (Xie, 2018), rmarkdown (Allaire et al., 2018), codebook (Arslan, 2019), and kableExtra (Zhu, 2018). All R packages used, including version information and dependencies, are documented in a lock file on OSF.

As the biases of interest are all related to measurement reactivity, we decided to report them on the original measurement scale, a Likert scale ranging from 0 to 4 (numbers were not shown to participants, who saw blank buttons). Incidentally, the results would have differed little had the effect sizes been standardized according to the sample variation: the overall SDs for all items were 0.95 to 1.18 (see Table 1) and the residual SDs in mixed models ranged from 0.85 to 1.10. The effect sizes we report are therefore approximately comparable with the Cohen's *d*s reported by Shrout et al. (2017). In most plots that follow, the Y axis ranges ±1SD from the mean to give a visual sense of the magnitudes of the effects relative to the sample variation.

We first investigated each of the randomized variables that were potentially related to measurement reactivity separately, inspecting graphs of the mean responses, response profiles, and reaction times. We then specified multilevel regression models, separately for each item. All models included a random intercept per woman to account for clustering by woman, thus adjusting for nonindependence in standard errors and individual differences in mean responses. We estimated all potential biases simultaneously. The models were adjusted for the number of days since the beginning of the diary as a set of 7 dummy variables, for the weekday as a set of six dummy variables (see also Supplementary Figure 1), and a dummy variable for the time period participants had been instructed to refer to at the beginning of the diary (the last 24 hours or for the time since the last entry if that was made less than 24 hours ago). As robustness checks, we additionally tested models that permitted biases to vary across women (random slopes), models that allowed biases to

have nonlinear effects, models restricted to participants who participated on every day in the first week, and models adjusted for the identity of the last item. After fitting the models, we tested whether residualizing for the estimated biases would non-negligibly affect our measures—that is, to what extent item scores residualized only for the covariates correlated with residuals from the full regression models. This approach allowed us to quantify whether a score without potential measurement reactivity bias was appreciably different[5] from the raw, unadjusted score.

We additionally tested a model to quantify how the number of randomly shown items on one day would predict likelihood of participation on the next day, i.e. to quantify the effect of randomly determined additional workload.

Results

We first visually inspected whether responses to items changed over time by inspecting the item means across days in the diary. No strong trend in mean levels was apparent (see Supplementary Figure 4). A plot of response times (see Figure 1) showed that participants responded more slowly on the first day than on later days. They took approximately 5s per item on day 1 before speeding up sharply, reaching an approximate asymptote at around 2–2.5s per item after about 30 days.

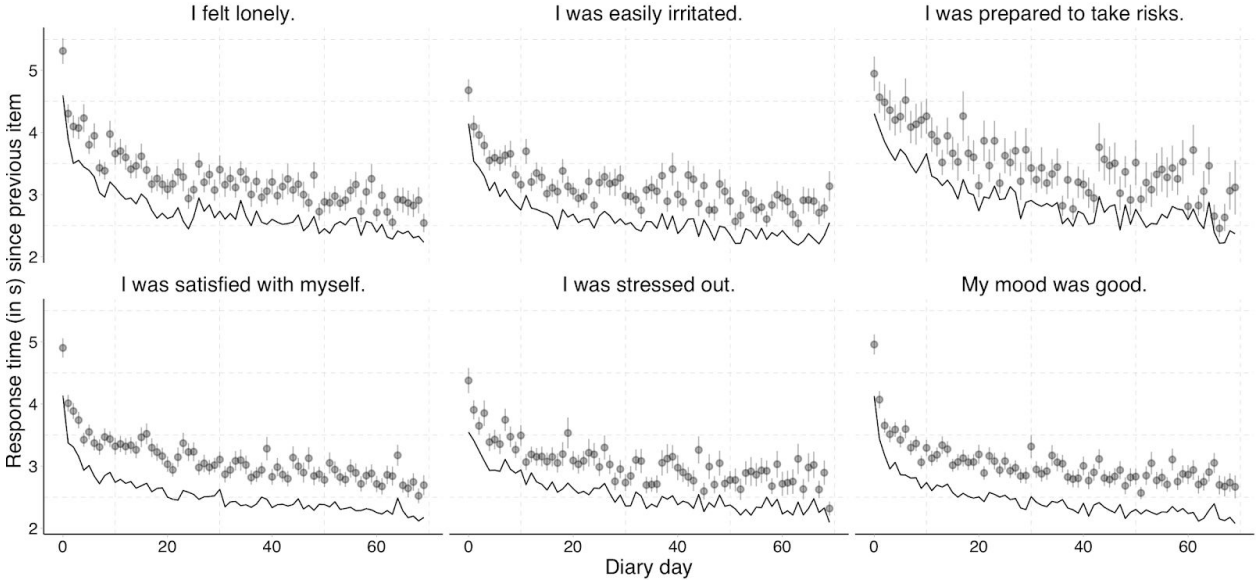

Figure 1. Mean response times per item (in s) by day in the diary. The black line shows the trimmed mean response time (with 10% of extreme values trimmed); the points show means with standard errors. Response times longer than 30s and responses given

[5] Of course, whether a score is "appreciably different" ultimately depends on the research question and the effect sizes being studied, but we think the correlation can provide some guidance.

out of order (negative times occurring when participants responded to items lower down the page before items higher up on the page) were excluded.

We then tested for initial elevation bias by grouping participants by the day they first saw each item. As Figure 2 shows, the first point of each colored line (representing groups who first saw each item on different days) is not consistently elevated above the long-term mean. This observation corresponds with the findings of our multilevel regression model, where point estimates for a dummy variable indicating the first time an item was shown ranged from −0.06 to 0.06 (Figure 6 and Supplementary Table 2) across items, with positive values reflecting initial elevation.

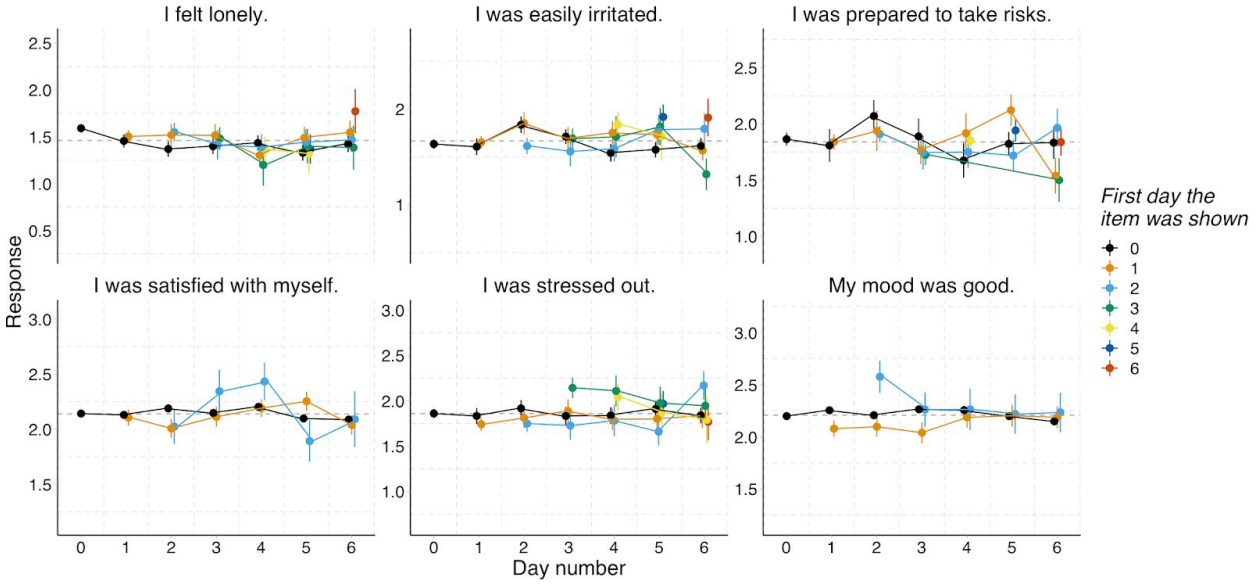

Figure 2. Mean response to each item by the number of days since starting the diary for groups sorted by the day on which they first saw each item (represented by colored lines). To make the magnitude of fluctuation from the mean visible, the Y axis scale is displayed from each item's mean ± 1 *SD*; values could range from 0 to 4. To reduce noise, we only show lines with at least 20 participants. Therefore, fewer lines are shown for items that had a higher probability of being shown each day. Findings showing that the initial point of each colored line exceeds the mean of the other lines on the day would be evidence for an initial elevation bias. The standard errors for the means do not account for the person-level structure of the data.

As Figure 3 shows, differences in item means depending on the order of item presentation were minute and inconsistent. In our multilevel regression model, point estimates of a linear variable for item order ranged from −0.03 to 0.02. Given the large

sample size, 95% confidence intervals for three items ("I was lonely", “I was prepared to take risks” and “I was satisfied with myself.”) excluded zero, but all confidence intervals remained within ±0.1 Likert points (see Figure 6). As shown in Figure 4, item means were not strongly associated with the identity of the preceding item either (see Supplementary Figure 19 for point estimates).

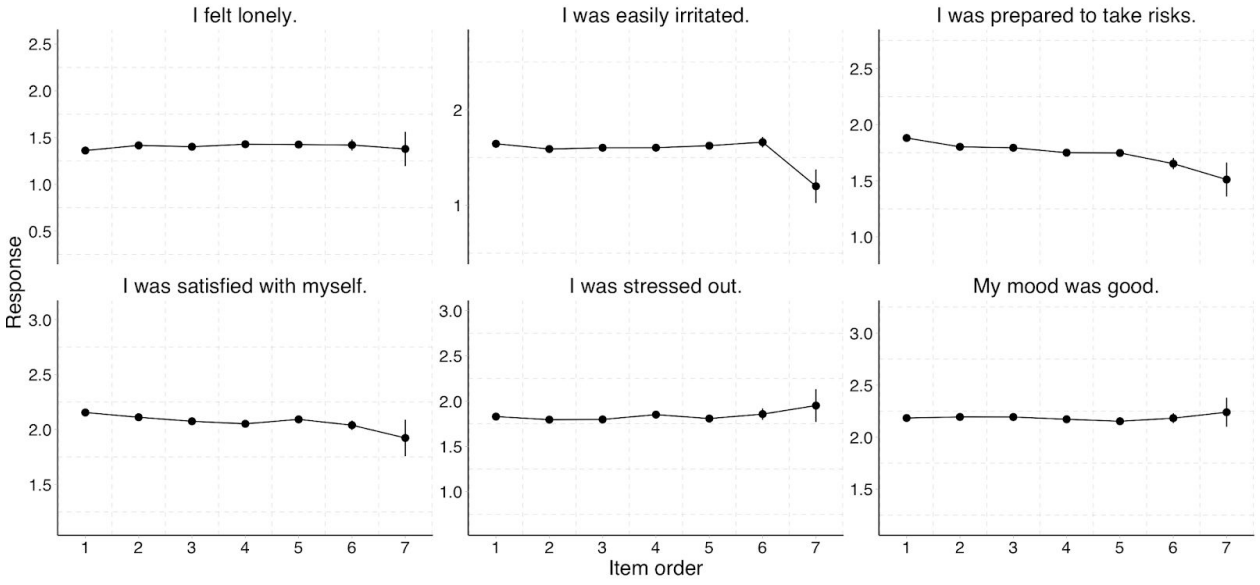

Figure 3. Mean response (± 1 *SE*) by item order. These raw estimates are still confounded with the number of items shown on that day and with relationship status (women in a relationship saw one additional item). The Y axis scale is displayed from each item’s mean ± 1 *SD*; values could range from 0 to 4. The standard errors are only visible for the sixth and seventh position because they were narrow. They do not account for the person-level structure of the data.

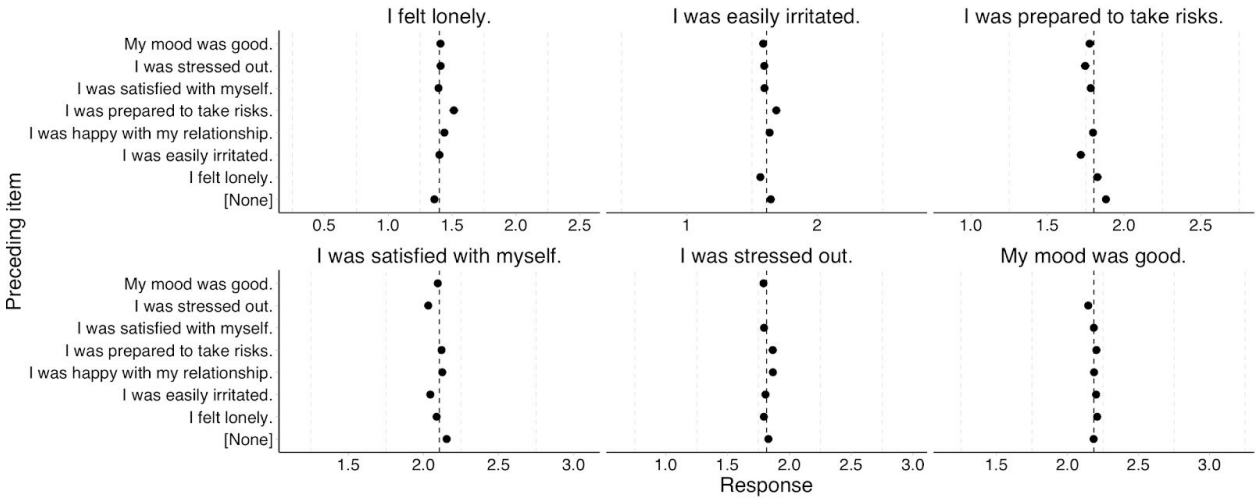

Figure 4. Average response (± 1 *SE*) by preceding item. Items preceded by no other item were necessarily also those shown first. The Y axis scale is displayed from each item's mean ± 1 *SD*; values ranged from 0 to 4. The standard errors for the means do not account for the person-level structure of the data.

We found only negligible effects of the number of items shown on item means either. In our multilevel regression model, point estimates of a linear effect of number of items shown were 0.00 to 0.01 across items (see Figure 6), once we had adjusted for item order. In other words, regardless of whether few or many items were shown on a day, the mean responses to each item were largely unchanged. As Figure 5 shows, response profiles were also largely unaffected by the number of items shown: people did not respond more extremely or tend more to the middle when more items were shown. Visually, the response profiles were very similar in shape to those plotted according to item order and item presented first (see Supplementary Figures 7, 11, and 16).

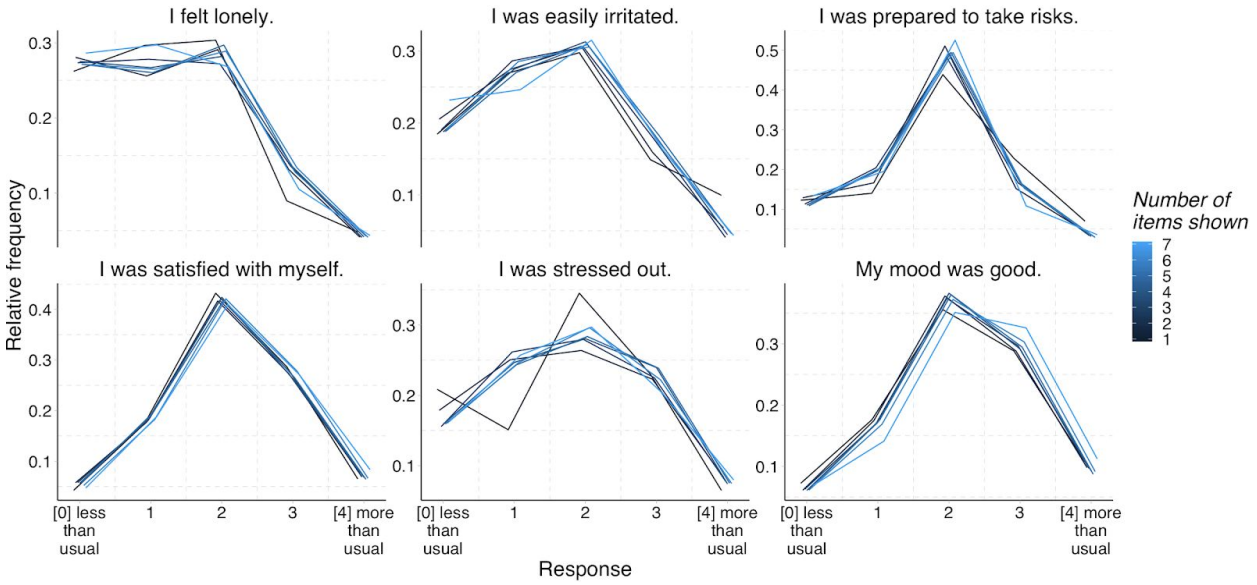

Figure 5. Response profiles plotted according to the number of items shown on that day.

We extracted the residuals after fitting our multilevel regression models and correlated them with the residuals from models only adjusted for covariates. The correlations were above .99 for all items. Our robustness analyses yielded substantially unaltered conclusions (see Supplementary Section 3.1). Results were similar for the three alternative sets of items (see Supplementary Section 4).

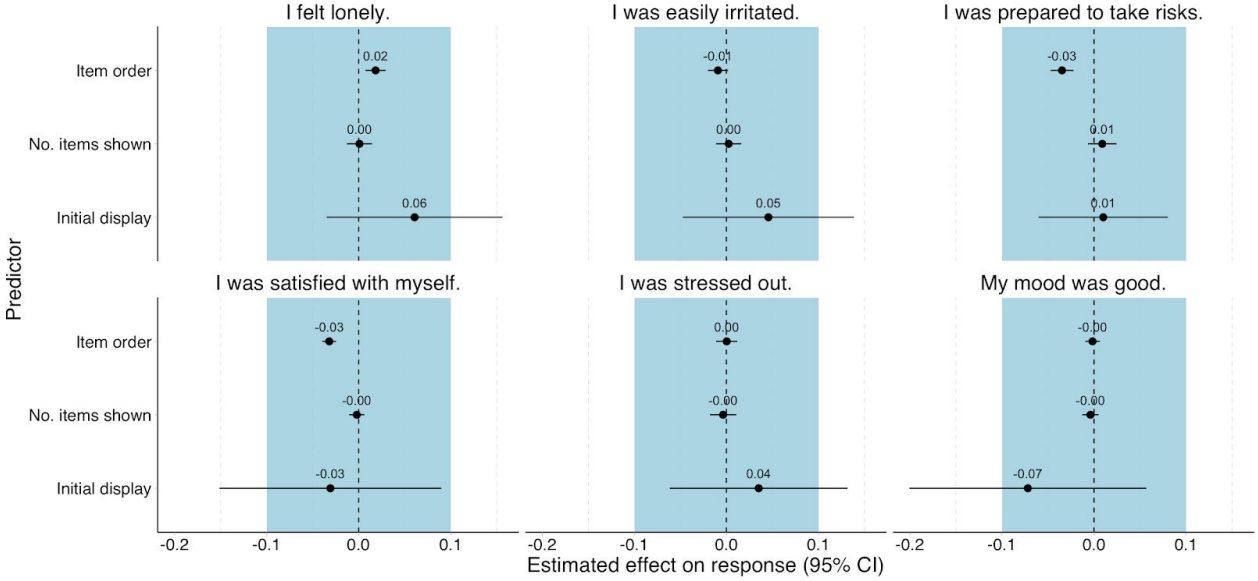

Figure 6. Coefficient plot showing the effects and 95% confidence intervals of the randomized variables on responses as estimated in a multilevel regression with person-level intercepts with covariates. The shaded area demarcates effects ±0.1 of a Likert point.

Whereas the mean response levels seemed to be minimally affected by the variables investigated here, response times clearly were affected. We fitted a multilevel regression model with person-level intercepts and predicted response time. Response times longer than 30 s and responses given out of order (negative times occurring when people responded to items further down the page before items higher on the page) were excluded. We found that responses were quicker on later days (Figure 1), the further down the page an item was positioned, and when an item had been answered before (Figure 7).

Finally, although participants were slightly more likely not to respond on the next day, if they had just completed a larger, randomly determined number of items, this difference was not significant (see Supplementary Figure 26 and Table 18).

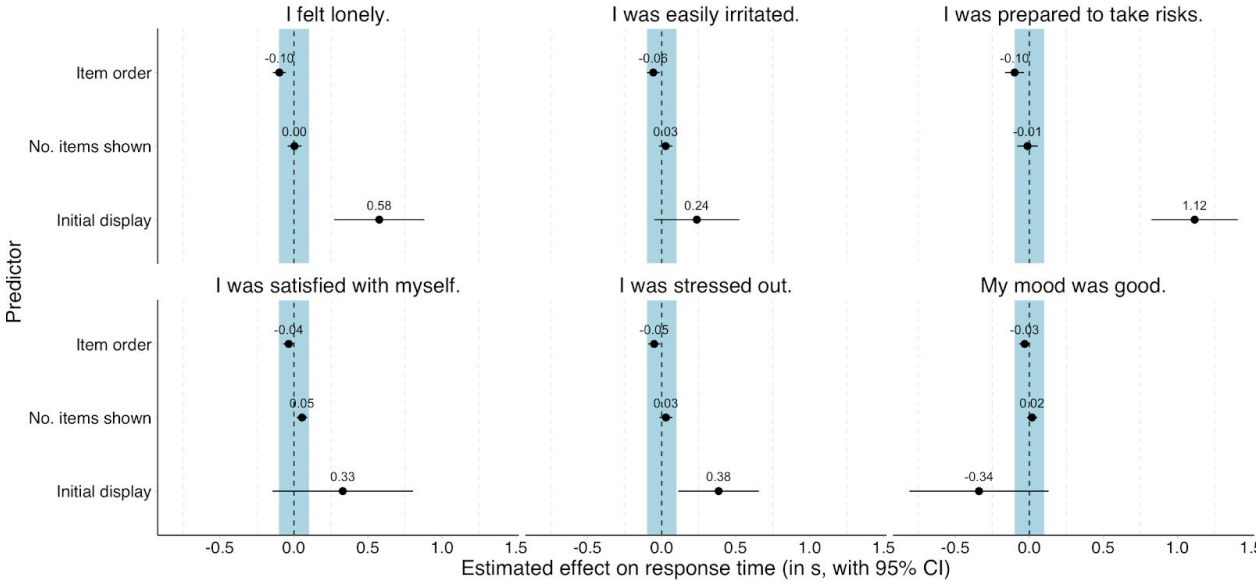

Figure 7. Coefficient plot showing the effects and 95% confidence intervals of the randomized variables on response time as estimated in a multilevel regression with person-level intercepts with covariates. The shaded area demarcates effects ±0.1 s.

Discussion

We employed a planned missingness design in a daily diary study to estimate and adjust for biases related to measurement reactivity. The estimated effects of the first day of item presentation, or initial elevation bias, were small, although potentially non-negligible effect sizes were still within the 95% confidence intervals. Our estimates were smaller (point estimates of −0.06 to 0.06 on a Likert scale with SDs from .95 to 1.18) than those reported by Shrout et al. (2017), where median estimates of Cohen's *d* ranged from 0.16–0.34 across studies. Also, our estimates were not consistently positive. Several reasons could explain this difference in results. Initial elevation bias might occur mainly on the survey level or only when a large number of related items is answered. The bias may be smaller for our items than for those administered by Shrout et al. Only three of our items—stress, loneliness and irritability—assessed negative mood states, for which Shrout et al. reported the largest

effects. Other potential explanations for the differences in the findings may be the sample (ours was restricted to German-speaking women) or the assessment procedure (e.g., we used blank Likert-type buttons without numeric anchors). Finally, Shrout et al.'s initial estimates of the initial elevation bias could themselves be elevated by another type of bias, namely, the "winner's curse" (Ioannidis & Trikalinos, 2005). This special case of regression to the mean happens when larger and significant effects are more likely to be noticed, written up, or published. If the initial elevation bias is smaller than first reported or can be reduced by making changes to the assessment procedure, this would be good news for the many social scientists who use subjective reports in their work.

Nevertheless, further work is needed to investigate the possibility of bias at the survey level. For example, our correlational results seem to indicate that loneliness was higher when people signed up for our study. However, this could be a real difference unrelated to measurement, such as a selection effect in which women who felt lonely were more likely to sign up for our study, or a treatment effect in which participation in the diary study truly reduced feelings of loneliness. Further experimental evidence is needed. The approach discussed here can be easily be combined with randomizing the onset (as in Shrout et al.) and frequency of repeated measures as well to distinguish, for example, a therapeutic effect, which could have a dose-response relationship with frequency of participation, from regression to the mean, which should be unrelated to measurement.

Other biases related to measurement reactivity that have been discussed in the literature include biases related to item order and carryover effects (Schimmack & Oishi, 2005). In our case, all biases we could quantify were negligible for the substantive questions we planned to address with the data, that is item order effects were small and participants were not much more likely not to respond if the workload in terms of items was higher. Still, at this early stage of research on measurement issues with repeated measures, we would caution researchers not to assume that our results will generalize to substantially different procedures or different populations. Until generalizable insights on measurement reactivity have been aggregated in the literature, researchers have two good reasons to employ planned missingness designs. First, they reduce participant burden while maintaining construct breadth (Revelle et al., 2016), and second, they can be used to estimate and adjust for measurement reactivity biases, as demonstrated here. Only if measurement reactivity can indeed lead to substantial bias, should researchers consider dropping initial data points. If reactivity is low, dropping initial data points would be an expensive countermeasure against bias and would lump together regression to the mean, treatment effects, and measurement reactivity, which may not be desirable depending on the research question. When data have already been collected and there is correlational evidence for initial elevation, dropping the first day of data may serve as a justifiable brute force robustness check, but it could also lead to an underestimation of treatment effects, so both analyses need to be reported and discussed. In new studies, we recommend that researchers do not assume problematic biases, but instead estimate and adjust them after randomization (this may include randomizing participants to an extensive pre-training condition), so that more data on bias generality accumulate.

Our proposed strategy—planned missingness—also has some costs, however. Unsurprisingly, given the name of the procedure, most downsides relate to the higher need

for planning. Can the statistical packages and models you plan to use cope with missing data?[6] Could validity issues arise if you do not ask the same questions everyday, for example because participants will not be sure whether to answer for the time frame since the last survey or since the last time the specific question was asked (Vogelsmeier et al., 2019)?[7] Fortunately, missingness resulting from randomization is generally ignorable in analyses; missing cases can simply be dropped without the need for multiple imputation or similar procedures. Nevertheless, researchers need to account for multiplicative missingness when planning studies. To give an example from our study, if we had been interested in examining whether there is a cross-lagged effect of mood on willingness to take risks, we would only have been able to include ~3% of the days in the diary without imputation: days where willingness to take risks was measured on two consecutive days and mood was measured on the previous day (i.e., 20%*20%*80%). Researchers should keep this multiplicative missingness in mind and assign sufficiently high probabilities to central items, or ensure that a central construct is tapped by multiple items to ensure coverage on all days (as we did for sexual desire, a central construct in our study), or incorporate missingness and imputation in appropriate simulations to ensure adequate sensitivity given the design parameters. Generally, simulating realistic data prior to data collection helps ensure that the planned models are estimable using the statistical packages you want to employ. For a given study, this will often depend on the research question, for instance, whether between- or within-person associations need to be estimable (Kirtley et al., 2020). While the effort of simulating data may provoke unease, it can foster an improved understanding of complex models (DeBruine & Barr, 2019), and help take some of the regret out of a preregistration experience. With experience sampling in general and even more so with planned missingness, researchers will often have to leave behind rules of thumb about sample sizes at different levels, out-of-the-box power analysis and similar simple tools (Kirtley et al., 2020; van Roekel et al., 2019). Here, we hope that increased planning mainly shifts the effort ahead in the schedule, not unlike Registered Reports (Chambers et al., 2015).

Overall, we are confident that these planning costs are worth paying, if they make it possible to answer questions about measurement reactivity, a concern affecting much of psychological measurement. We also encourage researchers to routinely collect reaction time data for survey responses. V(van Roekel et al., 2019) report that many studies do not report time per item and survey, even though it is important to understand the workload imposed by a survey over time. Additionally, by examining reaction times, we can find items which are difficult to understand or to answer. Currently, most researchers seem to rely on their intuition or small estimated correlations with similar items to find hard-to-parse items, but this tendency may inadvertently penalise construct breadth in scale construction.

---

[6] For example, while the psych R package (Revelle, 2018) implementations of multilevel generalizability (Shrout & Lane, 2012) and intraclass correlations (Shrout & Fleiss, 1979) cope well with missing data, several other packages implement intraclass correlation solutions that expect complete data.

[7] For further practical guidance on planning and preregistering studies including repeated measurements, see van Roekel et al. (2019) and Kirtley et al. (2020). Given that van Roekel et. (2019) found that none of the ambulatory assessment studies they reviewed reported a power analysis, researchers may currently be driven more by feasibility concerns.

Recording reaction times with sub-millisecond precision in web-browser-based studies is now a mature technology and available in open source software such as lab.js (Henninger et al., 2019) and formr.org (Arslan et al., 2019). Researchers who are willing to contribute their randomised survey data for a collaborative analysis of measurement reactivity across items and designs should contact the corresponding author.

A good rule of thumb may be that any measurement-related issue on which team members disagree when planning a study is a candidate for randomization. Disagreements can then be resolved by data. Teams may thus need to overcome a certain degree of experimentation aversion (Meyer et al., 2019), but they should remind themselves that preferring to remain ignorant of the consequences of a design decision is not a reasonable strategy (Himmelstein et al., 2019). Some further candidates for randomization include item wording specifics, the order and number of response categories for multiple choice and Likert-type questions, and whether there are follow-up questions (i.e., whether participants can learn to reduce their workload by answering a question with "No"). Beyond the analyses conducted here, future work might additionally (a) estimate potential elevation biases at the survey level by randomizing start dates after recruitment to find out whether starting a new survey causes initial elevation bias (as in Shrout et al.), (b) investigate the impact of (randomized) measurement frequency and participant burden (workload) on response times, dropout, compliance, nonresponse rates, and data quality (Eisele et al., 2020; Little & Rhemtulla, 2013)[8], (c) use cognitive modelling to integrate changes in response times and answers within a coherent framework, and (d) estimate empirically and theoretically grounded exclusion thresholds for overly fast responses, dependent on normative curves for the amount of experience participants have had with a survey and item (the current practice appears to be to use fixed thresholds).

In summary, to avoid flying blind with respect to the potentially deleterious effects of measurement reactivity, researchers using subjective reports should routinely randomize measurement frequency and order at the item and survey levels, whether the items are repeated or not, in clinical and population samples. Researchers who already use designs with planned missingness without randomization should consider adding randomization to reduce participant burden and also learn about its effects (Little & Rhemtulla, 2013). Adopting this approach can lead psychological research from a culture in which we estimate measurement reactivity correlationally, discuss it in footnotes, and hope for the best to one in which we randomize and estimate sources of measurement reactivity to adjust for and prevent these biases.

Author contributions

RCA, JCD, TMG, and LP planned and conducted the diary study. RCA analyzed the data and wrote the initial brief draft. JCD contributed to data cleaning and code checking. RCA and AKR wrote the extended draft. All authors critically revised the final manuscript.

[8] Eisele et al. (2020) found a small significant effect of number of items on noncompliance, which seems roughly consistent in size with our estimated non-significant small effect in Supplementary Figure 26.

References

Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., & Chang, W. (2018). *rmarkdown: dynamic documents for R*. https://CRAN.R-project.org/package=rmarkdown

Arslan, R. C. (2019). How to automatically document data with the codebook package to facilitate data re-use. *Advances in Methods and Practices in Psychological Science*, *2*(2), 169–187. https://doi.org/10.1177/2515245919838783

Arslan, R. C., Jünger, J., Gerlach, T. M., Ostner, J., & Penke, L. (2016). *Online diary study of ovulatory cycle shifts on sexual desire, sexual activity, and mating behaviour*. https://osf.io/d3avf/

Arslan, R. C., Walther, M., & Tata, C. (2019). formr: A study framework allowing for automated feedback generation and complex longitudinal experience sampling studies using R. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01236-y

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research*. Guilford Press.

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: realigning incentives in scientific publishing. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *66*, A1–A2. https://doi.org/10.1016/j.cortex.2015.03.022

Condon, D. M. (2018). The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. *PsyArXiv*. https://doi.org/10.31234/osf.io/sc4p9

DeBruine, L. M., & Barr, D. J. (2019). *Understanding mixed effects models through data simulation*. https://doi.org/10.31234/osf.io/xp5cy

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). *The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population*. https://doi.org/10.31234/osf.io/zf4nm

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. https://doi.org/10.1037/1082-989X.11.4.323

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2019). *lab.js: A free, open, online study builder*. https://doi.org/10.31234/osf.io/fqr49

Himmelstein, P. H., Woods, W. C., & Wright, A. G. C. (2019). A comparison of signal- and

event-contingent ambulatory assessment of interpersonal behavior and affect in social situations. *Psychological Assessment*, *31*(7), 952–960. https://doi.org/10.1037/pas0000718

Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, *58*(6), 543–549. https://doi.org/10.1016/j.jclinepi.2004.10.019

Kirtley, O. J., Lafit, G., Achterhof, R., Hiekkaranta, A. P., & Myin-Germeys, I. (2020). Making the black box transparent: A template and tutorial for (pre-)registration of studies using Experience Sampling Methods (ESM). In *PsyArXiv*. https://doi.org/10.31234/osf.io/seyq7

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

Knowles, E. S., & Condon, C. A. (1999). Why people say“ yes”: A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, *77*(2), 379. https://psycnet.apa.org/record/1999-03699-011

Little, T. D., & Rhemtulla, M. (2013). Planned Missing Data Designs for Developmental Researchers. *Child Development Perspectives*, *7*(4), 199–204. https://doi.org/10.1111/cdep.12043

Meyer, M. N., Heck, P. R., Holtzman, G. S., Anderson, S. M., Cai, W., Watts, D. J., & Chabris, C. F. (2019). Objecting to experiments that compare two unobjectionable policies or treatments. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(22), 10723–10728. https://doi.org/10.1073/pnas.1820701116

Palen, L.-A., Graham, J. W., Smith, E. A., Caldwell, L. L., Mathews, C., & Flisher, A. J. (2008). Rates of missing responses in personal digital assistant (PDA) versus paper assessments. *Evaluation Review*, *32*(3), 257–272. https://doi.org/10.1177/0193841X07307829

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2018). *psych: procedures for psychological, psychometric, and personality research* (Version 1.7.3) [Computer software]. https://CRAN.R-project.org/package=psych

Revelle, W., Condon, D. M., Wilt, J., French, J. A., Brown, A., & Elleman, L. G. (2016). Web and phone based data collection using planned missing designs. In N. G. Fielding, R. M. Lee, & G. Blank (Eds.), *The SAGE Handbook of Online Research Methods*. SAGE. http://mobile.personality-project.org/revelle/publications/websapa.final.pdf

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Schimmack, U., & Oishi, S. (2005). The influence of chronically and temporarily accessible information on life satisfaction judgments. *Journal of Personality and Social Psychology*, *89*(3), 395–406. https://doi.org/10.1037/0022-3514.89.3.395

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

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037//0033-2909.86.2.420

Shrout, P. E., & Lane, S. P. (2012). Psychometrics. In T. S. Conner & M. R. Mehl (Eds.), *Handbook of research methods for studying daily life* (pp. 302–320). Guilford Press. https://nyu.pure.elsevier.com/en/publications/psychometrics

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. (2017). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1712277115

Silvia, P. J., Kwapil, T. R., Walsh, M. A., & Myin-Germeys, I. (2014). Planned missing-data designs in experience-sampling research: Monte Carlo simulations of efficient designs for assessing within-person constructs. *Behavior Research Methods*, *46*(1), 41–54. https://doi.org/10.3758/s13428-013-0353-y

van Roekel, E., Keijsers, L., & Chung, J. M. (2019). A Review of Current Ambulatory Assessment Studies in Adolescent Samples and Practical Recommendations. *Journal of Research on Adolescence: The Official Journal of the Society for Research on Adolescence*, *29*(3), 560–577. https://doi.org/10.1111/jora.12471

Vogelsmeier, L. V. D. E., Vermunt, J. K., van Roekel, E., & De Roover, K. (2019). Latent Markov Factor Analysis for Exploring Measurement Model Changes in Time-Intensive Longitudinal Studies. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 557–575. https://doi.org/10.1080/10705511.2018.1554445

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software, Articles*, *59*(10), 1–23. https://doi.org/10.18637/jss.v059.i10

Xie, Y. (2018). *knitr: a general-purpose package for dynamic report generation in R*. https://CRAN.R-project.org/package=knitr

Zhu, H. (2018). *kableExtra: construct complex table with “kable” and pipe syntax*. https://CRAN.R-project.org/package=kableExtra