Routinely randomise the display and order of items 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 Goettingen
4 Leibniz ScienceCampus Primate Cognition

Corresponding author: Ruben C. Arslan (ruben.arslan@gmail.com)

Abstract: Recent psychological research has experienced a stark increase in the use of repeated subjective reports, such as online, smartphone-based daily diaries. This development holds great opportunities to study causal processes and developmental change, but it also brings new challenges. As is often the case in psychology, interest in specific substantive questions can outrun measurement research, so that many measurement and design decisions are made ad hoc and not evaluated rigorously. Recent work has added initial elevation bias to the list of common pitfalls that should be avoided when using subjective reports. Initial elevation bias refers to the phenomenon that negative states (i.e, thoughts, feelings, behaviors) in subjective reports are elevated when first assessed, as compared to later assessments. In this article, we employ a planned missingness design in a daily diary study of more than 1,200 individuals that were assessed over a period of up to 70 days to estimate and adjust for initial elevation bias. First, we found only a negligible bias related to initial presentation and item order: items were not answered differently depending on when and where they were shown. Second, we show that residualising these biases had minor effects. We conclude from our findings that the initial elevation bias may be more limited than previously reported and may only act at the level of the survey, not at the item level. We encourage researchers to make design choices that will allow them to routinely ascertain potential measurement reactivity biases empirically in their studies. Specifically, we advocate that researchers should routinely randomise item display and order in planned missingness designs, so that they can estimate biases affecting subjective reports. Another benefit of routinely randomizing item display is that it allows constructing brief survey instruments without compromising the construct breadth and the number of constructs covered.

Introduction

Repeated subjective reports, in which individuals provide multiple reports in a short time span on thoughts, feelings, and behaviors, have been increasingly used in many disciplines including social, personality and health psychology. They have a wide range of applications, such as in research on individual development, well-being and health, and are often used to examine causal processes, developmental change and individual variability therein. A reason for the increased popularity of this type of data is that they became relatively easy to assess through the increased use of digital devices such as smartphones and wearables. Repeated subjective reports open up the possibility to understand daily life and causal processes as well as new possibilities to tailor interventions to individuals' unique needs (Bolger & Laurenceau, 2013).

As research increasingly relies on them, potential validity issues with repeated subjective reports were brought up. Instead of random measurement errors, response biases were discussed as one reason for the validity problems (Eric S. Knowles & Condon, 1999). In particular, researchers have reported that the reporting of the severity of negative states, such as anxiety and depression, decreased across repeated reports (E. S. Knowles, Coker, Scott, Cook, & Neville, 1996; Sharpe & Gilbert, 1998). Researcher thought this pattern reflected an "attenuation effect", in which later responses were biased downward because of measurement reactivity.

Recently, however, Shrout et al. ((Shrout et al., 2017) showed that instead of a decline in later reports, self-reports of subjective, negative states tend to be elevated when first assessed. The authors randomised subjects to different starting dates in diary and experience sampling studies in which participants repeatedly reported on their mood, thoughts, and behavior. They showed that initial reports were biased upward (initial elevation), even in groups that had been randomised to start after a delay following their enrolment. Before this study, such discrepancies between early and later reports were thought to be either the result of a later attenuation bias or a selection bias in which people are more likely to enrol in a study if they are, for example, particularly anxious and then regress to the mean. If, as suggested by Shrout et al. (2017), the discrepancies are instead due do an initial elevation bias, this is bad news for research using subjective reports, because this bias would affect even non-repeated self-reports, perhaps the most frequent method of data collection in social science. Such a bias could, among other

things, explain spurious symptom "improvements" in control groups, exacerbating better-known biases like regression to the mean and placebo response, because initial reports would overestimate the real value. It would also mean that cross-sectional studies of negative subjective states would always overestimate the mean compared to repeated panel studies of the same population. Given these wide-ranging consequences, Shrout et al. called for researchers to further investigate the initial elevation bias.

Shrout et al. stopped short of establishing whether the bias occurs at the item level or the survey level, that is whether people give elevated responses when they first start a new survey or when they see a new item for the first time. If the bias affects items, initial elevation bias would even affect newly introduced questions in panel studies and would be even harder to reduce. We therefore sought to investigate whether the bias occurs at item or survey level. Shrout et al. found that items about negative but not positive states were affected, which suggests that the bias occurs at the item level, but the authors also imply that starting the study may be a causal factor, which suggests the survey level. Only study starting dates were randomised in their studies. Based on this interpretation, Shrout et al. suggested two potential countermeasures against the bias: a) to drop initial observations or b) to familiarize subjects with survey. Needless to say these strategies are expensive and would also incur the loss of partially valid information.

We disagree with these expensive proposed countermeasures and instead recommend the design decision to randomise the supposed causes of bias when planning a study. While dropping observations may be a last resort as a robustness check when working with existing data, randomisation allows researchers to estimate and adjust for bias. It is not only less wasteful, but also compatible with another piece of widely ignored best practice advice, namely planned missingness designs (Condon, 2018; Revelle et al., 2016; Silvia, Kwapil, Walsh, & Myin-Germeys, 2014).

Reducing waste and redundancy in their studies is near and dear to many researchers who collect repeated subjective reports. They face the dilemma of wanting to keep their surveys brief in order to avoid drop-outs and 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 owing to item order, initial elevation, question familiarity, workload, and measurement reactivity can be addressed with one design decision

made before the data acquisition; namely by not showing all items on all days, but instead randomly selecting a subset in random order for each day. Doing so, surveys can be kept brief and expected biases can be estimated and, if non-negligible, statistically adjusted for. Missing values that result from randomising item display are ignorable (Rubin, 1976) and usually require no greater statistical expertise to handle than the analysis of multilevel data requires anyway. Changing questions and their order on a daily basis can also keep monotony at bay despite repetition, so that participants do not respond "on autopilot" and drop out less often (Silvia et al., 2014). As a result, a planned missingness design for repeated subjective reports may thus reduce systematic missingness by reducing participant fatigue and the contingency between responding to later items and fatigue (i.e., that participants are more fatigued and less motivated when responding to later items; (Palen et al., 2008).

In the present study, we re-used data from a diary study employing a planned missingness design to estimate initial elevation bias using a different method. We make the case that estimating and adjusting measurement reactivity related biases routinely, as demonstrated here, is cheap, easy, and desirable.

Methods

The present study was mainly designed to investigate ovulatory changes in subjective states, but it is also ideally suited to investigate initial elevation bias and other biases related to measurement reactivity, because for most items we randomised whether and in which order they were shown 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 (e.g., the online platform psytests.de, advertisement on okCupid.com and Facebook, and mass mailing lists of German-speaking university students) as well as direct invitations of suitable candidates who took 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 either direct

payment of participants with an amount ranging from 25€ up to 45€.[1] Alternatively, participants had the chance of winning prizes with a total value of 2,000€.[2] Students of the University of Goettingen were also able to earn course credit. For all three rewards, the amount of credit, money, or lots depended on the regularity of participation. At the end of the study, every participant received a personalised graphical feedback as a further incentive.

Study structure

Women participated in an online study named “Alltag und Sexualität [Daily Life and Sexuality]” implemented using the survey framework formr.org (Arslan, Walther, & Tata, in press). The study was introduced as an online diary which aimed to examine the interaction of sexuality, psychological well-being, experience of romantic relationships, and everyday experiences. The study had six main stages, but we will focus on the repeated diary in this study. After consent forms, participants filled out a demographic questionnaire, which was used for an initial screening phase for suitable participants. After participants were informed whether they would be paid or participate in the lottery, a personality questionnaire followed, which was, irrelevant for the current study as it was a single assessment. A day after these surveys, women started the online diary. Over a period of 70 days, women received an online invitation via email at 5 pm (they received text message reminders if they had entered their mobile phone number and missed several diary invitations). The online diary could be filled out until 3 am on the following day and included questions about their mood, daily activities, and questions concerning their sexuality. Items were randomised within grouped blocks of varying size. The items most central to the main questions of the planned study were shown every day while items of lower importance randomly appeared 20-80% of the time. After the diary had ended, the fourth step was a social network questionnaire and a final follow-up questionnaire that assessed whether important changes occurred during the diary (both of which are irrelevant for the current study).

[1] Only women fulfilling certain sample criteria are offered direct payment. Those were being under the age of 50, being heterosexual, having a regular menstruation and being pre-menopausal as well as having not taken any hormonal or psychoactive medication and no hormonal contraception in the last three months. Additionally, women were only paid if they were not trying to get pregnant or had been pregnant and/or breastfeeding within the last three months.

[2] The prizes of the lottery included an iPhone, an iPad and forty 20€ Amazon coupons.

Data subset used for the present study

We used 57,061 daily diary entries (mainly closed-ended questions, which this investigation focuses on), reported by 1,259 women over up to 71 days (*M* = 45 days, *SD* = 22). Women were between 18 and 61 years of age (*M* = 26.7, *SD* = 7.3) and had on average 15.2 years of education (*SD* = 4.8). Two thirds (66%) were students, 31% were employed. Ten percent were married, two percent engaged, 50% were in a committed relationship, and 31% were single, with the remainder being in non-committed relationships. Twelve percent had children. Subjects gave their informed consent (survey studies are exempt from ethics committee approval under German regulations). Six questions about stress, loneliness, mood, risk taking, self-esteem, and irritability were presented on the first page of the online diary and are the focus of this investigation (see Table 1). Each day, a random subset of these items was shown as the first items on the first page, in random order. We randomised item selection and order on a daily basis and separately for each participant. Items were first shuffled, then a pseudorandom number was drawn from a uniform distribution to determine whether each item would be shown. This procedure was implemented in formr.org using R (Arslan et al., in press). Participants could respond to each item on a 5-point Likert scale that ranged from “less than usual” to “more than usual”. Pole labels were placed left and right of five blank, equally-sized buttons.

Because of our planned missingness design with randomised display and order, the following variables were randomised: the number of times an item was seen previously (conditional on adjusting for day number in the diary), the display order, the number of items shown on that day, and the identity of the preceding item(s).

Table 1. The items investigated here (wording translated to English from German). Because of the randomisation, sample sizes differ by item. The percentage reflects the probability that an item was shown on each day.

Item	N[women]	N[days]	Days/Woman	Mean	SD	% shown
My mood was good.	1250	45534	36	2.19	1.02	80%
I was easily irritated.	1209	22763	19	1.61	1.12	40%
I felt lonely.	1214	22695	19	1.40	1.14	40%
I was prepared to take risks.	1180	11364	10	1.79	0.95	20%
I was satisfied with myself.	1248	45545	36	2.11	0.97	80%
I was stressed out.	1224	22736	19	1.81	1.18	40%

Note. N[women] = the number of women who have seen the item at least once. N[days] = the number of days the item was shown. Days/Woman = The number of days a woman has on average seen the item. The percentage shown reflects the nominal and empirical percentage of the 57,061 days in the diary on which the item could be shown.

We selected these items presented in Table 1 because such general mood and state items are widely used in psychological research. In order to conduct a robustness check, we also examined some of the other items in the diary that were more specific to the research questions of the study. We chose three item sets that differed from the items on the first page. We used five items on how participants spent their time that were also answered on a response scale from "less than usual" to "more than usual", because these items were not about subjective states. To vary the response scales, we used six items on partnered sexual desire, which were answered on a response scale from "inaccurate" to "very accurate" and three items on partner jealousy that were answered on a response scale from "not at all" to "very much". The items are described in more detail in the online supplement (https://rubenarslan.github.io/initial_elevation_bias/).

Analysis

The biases of interest are all related to measurement reactivity. Thus, we decided to report them on the original measurement scale, a Likert scale from 0 to 4. Incidentally, standardising the effect sizes according to the sample variation would not have led to very different results. The overall SDs for all items were around 1 (see Table 1) and the residual SDs in mixed models ranged from 0.85 to 1.10. Reported effect sizes are therefore approximately comparable to the Cohen's *d*s reported by (Shrout et al., 2017). In all plots that follow, the Y axis ranges from the mean ±1SD to give a visual sense of magnitudes of the effects relative to the sample variation.

We first investigated each of the randomised variables that were potentially related to measurement reactivity separately using graphs of the mean responses, response profiles, and reaction times. We then specified several multilevel models, separately for each item. All models included a random intercept per woman to account for clustering by woman to account for non-independence in the standard errors for the effect and individual differences in the mean responses. We estimated all potential biases simultaneously. For each item, we tested both a model including only non-varying biases (the same magnitude for each participant) and a model that permitted biases to vary across woman (random slopes). After fitting the models, we tested whether residualising for the estimated biases would non-negligibly affect our measures, that is how high raw item scores correlated with residuals from the regression models. This approach allowed us to see whether a score without potential measurement activity bias was appreciably different from the raw, uncorrected score.

Because of the randomisation, selection should not play a role in these analyses. However, as is common in longitudinal studies, including the Shrout et al. study and our own, there is incomplete data. If dissatisfied individuals are more likely to discontinue the study, we might also see an initial elevation in dissatisfaction. To estimate such drop out effects, we tested the initial elevation bias both only for participants who did not miss a day during the first week and for all participants, including those who missed days.

Results

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

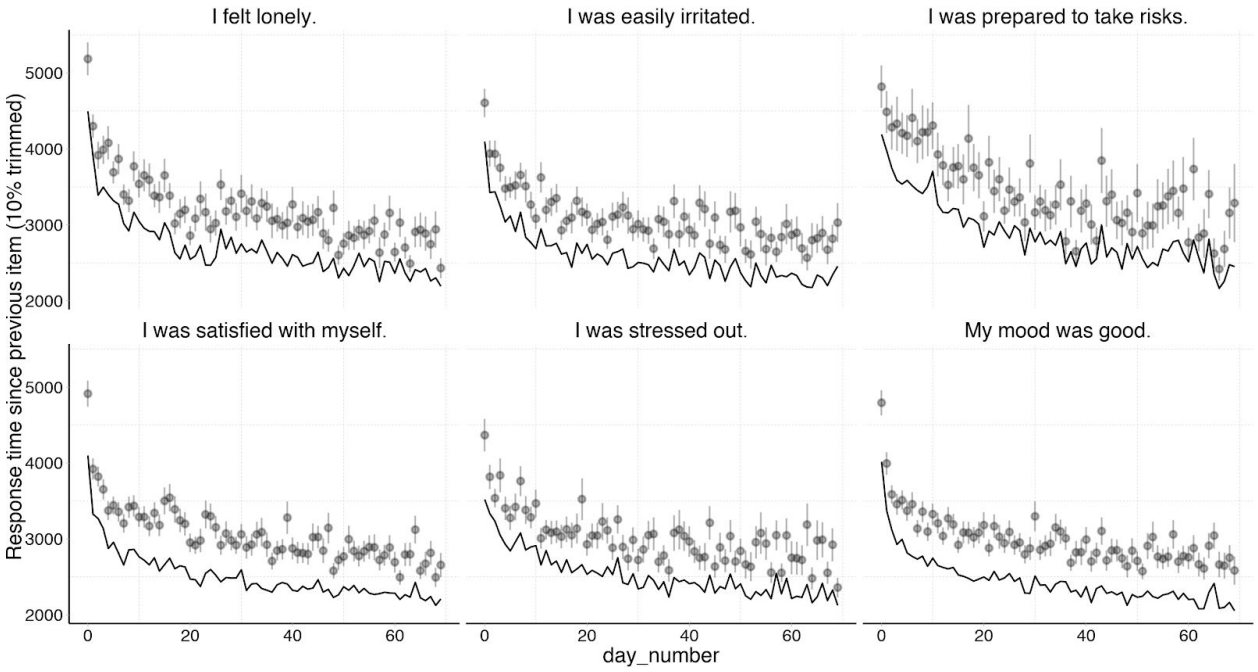

Figure 1. The Y axis shows the response time in milliseconds relative to the previous item. The X axis shows the day in the diary. The black line shows the trimmed mean response time (with 10% of extreme values trimmed), the points show means and standard errors. Responses taking longer than 30 seconds and responses out of order (items lower on the page answered before items higher on the page) were excluded. The standard errors for the means do not account for the person-level structure of the data.

We then tested for an initial elevation bias. In Figure 2, we grouped participants by the first day they saw each item. As one can see in Figure 2, the first point of each coloured line is not consistently elevated above the long-term mean. The visual inspection corresponds with the findings, as point estimates for a dummy variable indicating the first time an item was shown ranged from -0.06 to 0.07 (SEs ranged from 0.04 to 0.07) across items, where positive values reflect initial elevation.

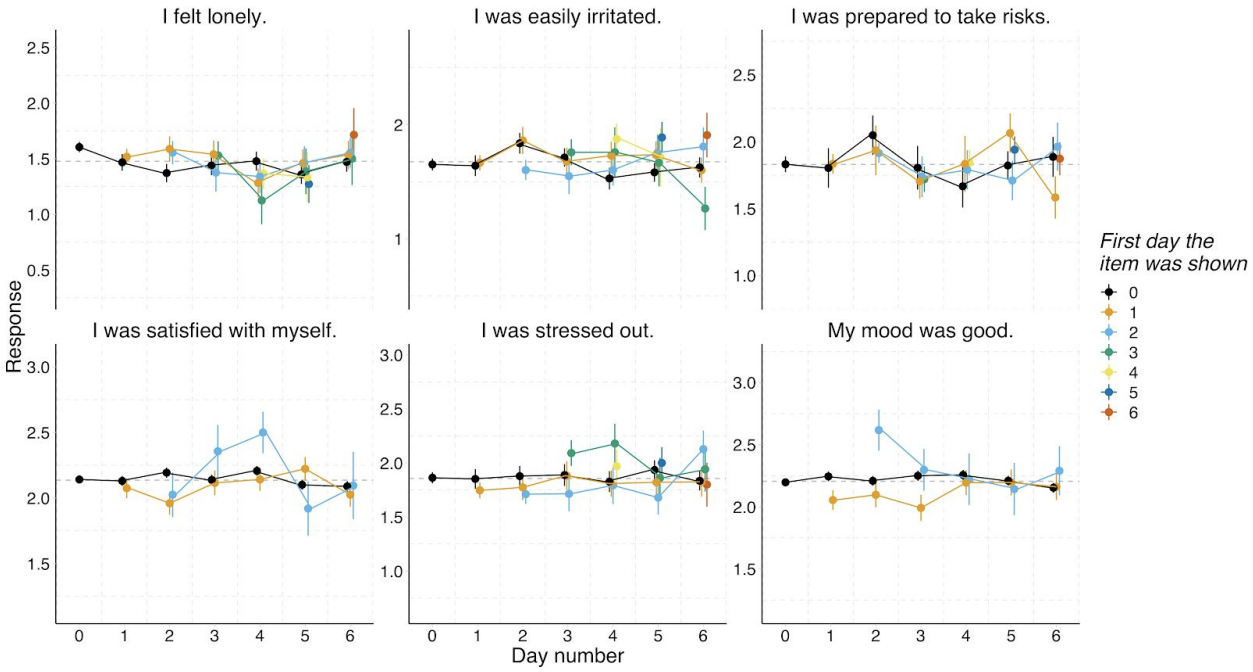

Figure 2. The figure shows the mean response to each item on the Y axis according to the number of days since starting the diary on the X axis. The Y axis scale is displayed from each item's mean ± 1 SD to make the magnitude of fluctuation from the mean visible, values could range from 0 to 4. Different-coloured lines reflect different starting days (the day of the diary we first asked the item). We only show lines based on at least twenty participants to reduce noise. Therefore, fewer lines are shown for items with a higher probability of being shown each day. Wherever the initial point of each coloured line exceeds the mean of the other lines on the day, this would be evidence for initial elevation bias. The standard errors for the means do not account for the person-level structure of the data.

As we show in Figure 3, there were only minute and inconsistent differences in item means according to item order. Given the large sample size, the small effects of later question order were significant for the item "I was prepared to take risks" and for the item "I was satisfied with myself". Point estimates of a linear variable for item order ranged from 0 to -0.04 (SEs were ≤ 0.01). As we show in Figure 4, the identity of the immediately preceding item was also not strongly associated with the mean levels.

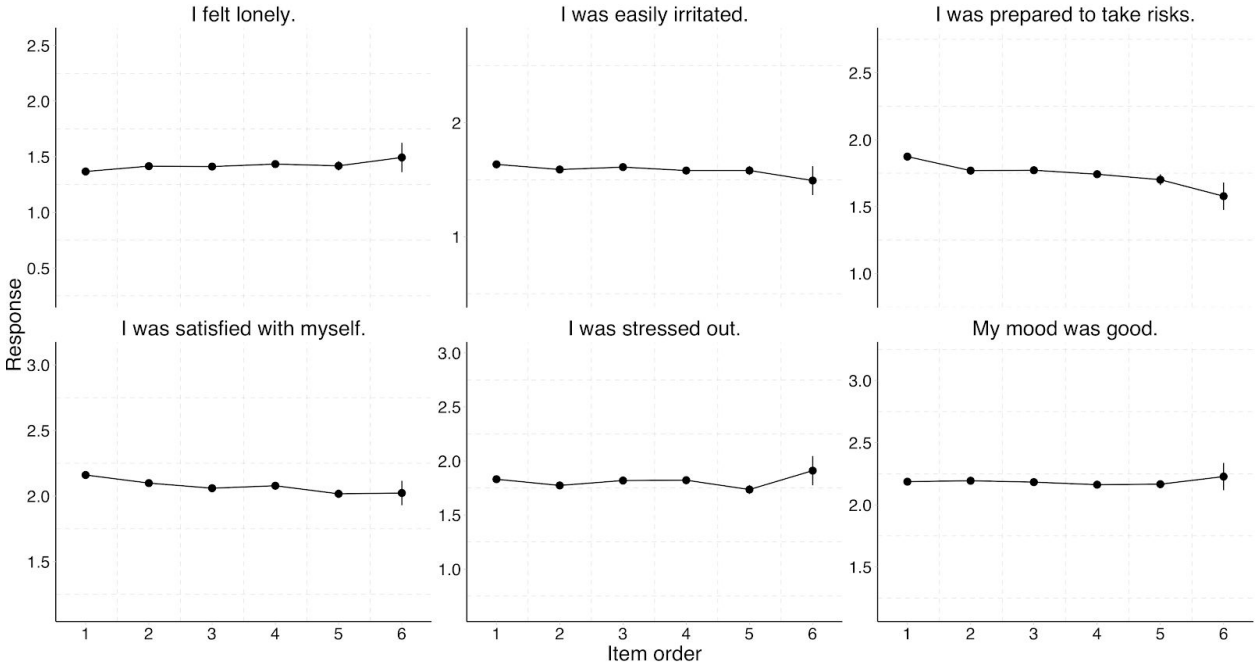

Figure 3. Average response (± 1 SE) according to item order. These raw estimates are still confounded with number of items shown on that day. The Y axis scale is displayed from each item's mean ± 1 SD (value range from 0 to 4). The standard errors are only visible for the sixth position, because they are narrow. The standard errors for the means do not account for the person-level structure of the data.

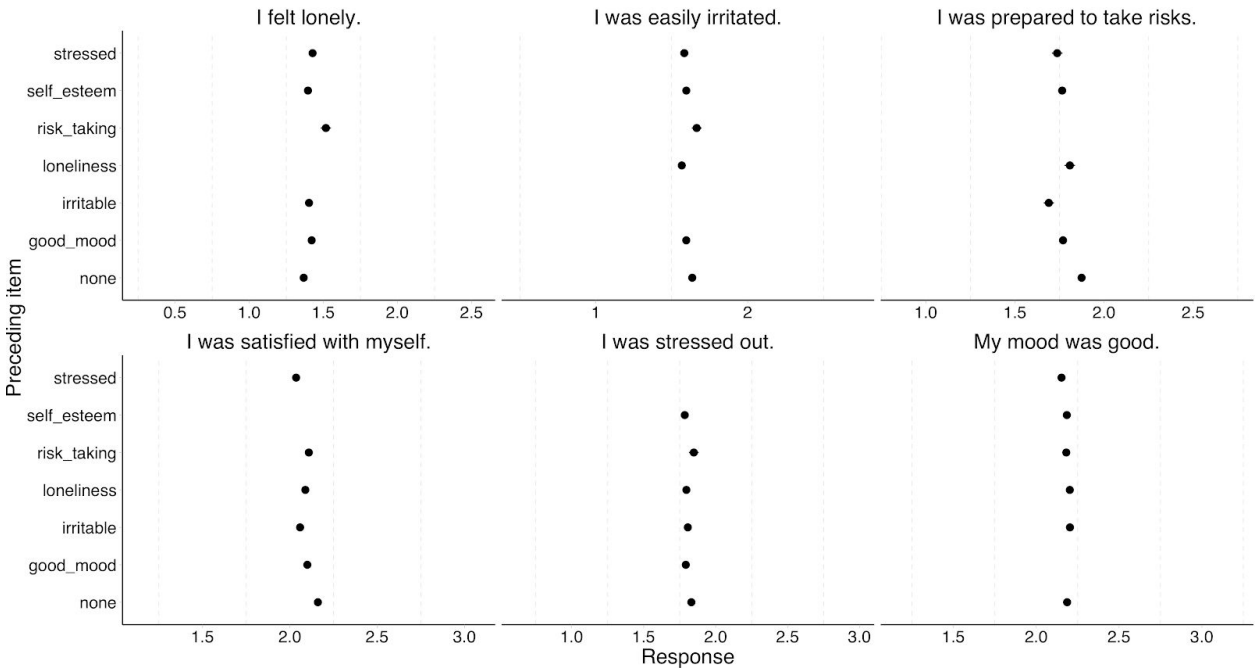

Figure 4. Average response (± 1 SE) according to the preceding item. Items preceded by no other item are necessarily also those shown first. The Y axis scale is displayed from each item's mean ± 1 SD (value range from 0 to 4). The standard errors for the means do not account for the person-level structure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

As we show in Figure 5, we found no significant effect of the number of items shown. Point estimates of a linear effect of number of items shown were -0.01 to 0.01 across items (SEs were ≤ 0.01), once the item order was adjusted for. This means that regardless of whether few or many items were shown on that day the mean responses to each item were largely unchanged.

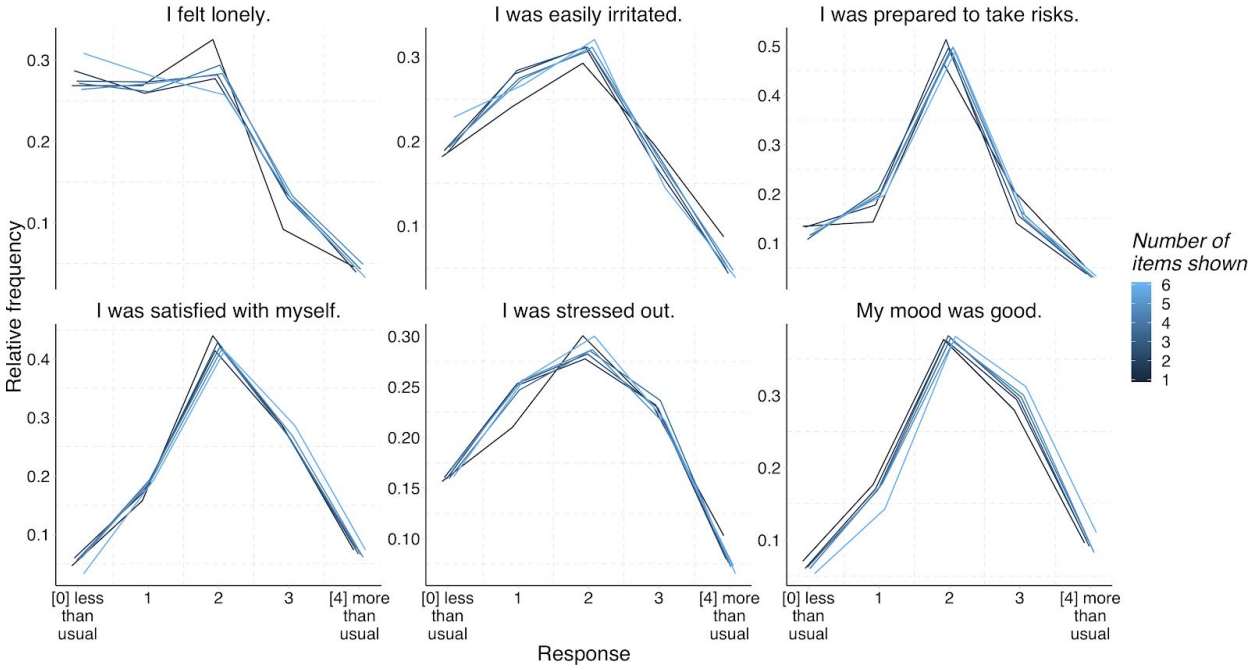

Figure 5. Visually, response profiles were very similar in shape when plotted according to the number of items shown on that day.

We transformed all aforementioned randomised variables into dummy factor variables to permit nonlinear effects. In a multilevel model, we estimated the effects of all randomised variables jointly. We additionally adjusted for the number of the day (how many days since the diary began), and the reference time period (participants were instructed to refer to the time since the last entry if it was less than 24 hours ago or else to the last 24 hours). Model coefficients are reported in the online supplement (https://rubenarslan.github.io/initial_elevation_bias/). Next, we regressed all aforementioned variables on each item in simple multiple linear regressions. We extracted the residuals after fitting the models and correlated them with the raw item scores. The correlations were above .99 for all items. Based on visual analysis, response profiles were similarly unaffected (see Figure 4 and online supplement). We also estimated a multilevel model in which the effects of the randomised variables were linear but allowed to vary across woman (i.e., varying slopes). Again, the residuals were correlated with scores residualised only for

person-level intercepts above .99. We estimated correlations between raw scores and residuals from a simple linear regression for three further sets of items (five items on time use, six item on partnered sexual desire, and three items on partner jealousy). In all cases, item raw scores correlated with the residuals at above .99.

Whereas mean levels of responses seemed to be minimally affected by the variables investigated here, response times clearly changed. We fitted a multilevel regression model with person-level intercepts and predicted response time relative to the response to the previous item. Response times longer than 30 seconds and responses out of order (artificially occurring negative times when people responded to items further down before items higher up on the page) were omitted. We found that on later days and when an item is further down on the page, responses were quicker (Figure 5). There was little evidence that the number of items and the times a specific item was shown had additional effects.

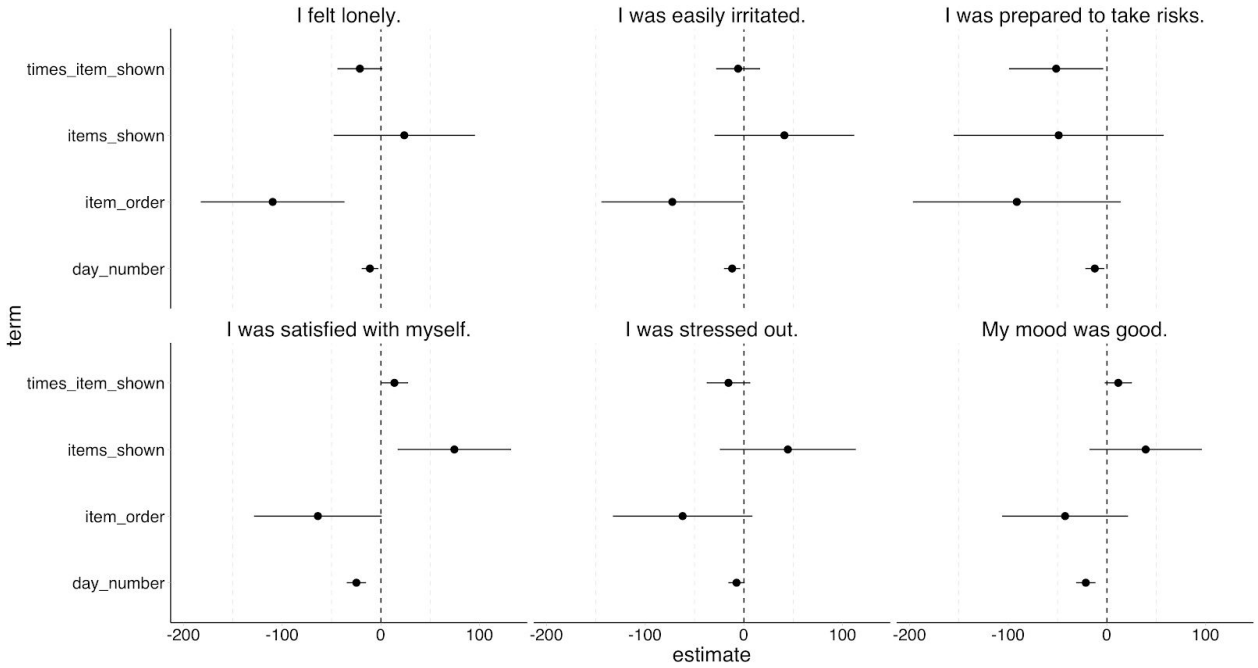

Figure 6. Coefficient plot showing regression estimates of response time and 99% confidence intervals based on a multilevel regression with person-level intercepts.

Our analyses and the code used to produce all figures and tables are fully documented online (Arslan, 2018).

Discussion

The goal of this study was to employ a planned missingness design in a daily diary study in order to estimate and adjust for initial elevation bias. We estimated negligible effects of the first day of item presentation. Our estimates were far smaller and not consistently positive (point estimates of -0.05 to 0.06 on a Likert scale with SDs from .95 to 1.14) compared to the initial elevation bias that Shrout et al. reported (median estimates of Cohen's *d* ranged from *d* = 0.16-0.34 across studies). This may be the case because the initial elevation bias occurs on the survey level or because the bias is smaller for our items (only two of our items, 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 (our sample included only German-speaking women) or the assessment procedure (e.g., we used blank Likert-typed buttons without numeric anchors). Ironically, Shrout et al.'s initial estimates of the initial elevation bias could themselves be elevated by another bias, specifically the "winner's curse" or the Proteus phenomenon (Ioannidis & Trikalinos, 2005). This special case of regression to the mean happens if publication is more likely for large and significant effects. If the initial elevation bias is smaller than initially reported or can be reduced using changes to the assessment procedure, this would be good news for the many social scientists who use subjective reports in their work.

Still, further work is needed to investigate the possibility of bias at the survey level. Our correlational results for loneliness 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 reduced feelings of loneliness. Further experimental evidence is needed. The approach discussed here can be easily extended to randomising the onset of repeated studies as well.

Other biases related to measurement reactivity that have been discussed in the literature are biases related to item order and carryover effects (Schimmack & Oishi, 2005). We found that in our case, all biases we could quantify were negligible for the substantive questions we plan to answer with the data. Still, at this early stage of research on measurement issues with repeated measures, we would caution researchers not to assume our results will generalize when using

substantially different procedures and different participant populations. Until generalizable insights on measurement reactivity have been aggregated in the literature, we think researchers have a second reason to employ planned missingness designs. Not only can they use them to reduce participant burden while maintaining construct breadth, they can also estimate measurement reactivity biases and adjust for them. Contra Shrout et al., we think dropping initial estimates or training participants with survey instruments are wasteful countermeasures against bias. Only when data has already been collected may dropping the first day of data become a justifiable bruteforce robustness check. In most other cases, we recommend not to assume problematic biases, but instead to estimate them after randomisation.

Our alternative proposition, planned missingness, however, has some costs, too. Unsurprisingly, given the name of the procedure, these are mostly related to a higher need for planning: Missingness resulting from randomisation is ignorable in analyses, so that missing cases can simply be dropped without need for multiple imputation or similar procedures. Still, researchers need to account for multiplicative missingness when planning their study. In our case, if we had been interested in examining whether there is a cross-lagged effect of mood on risk taking, for example, we would only be able to include ~3% of the days in the diary, because we would need days where risk taking was measured on two consecutive days, and mood was measured on the previous day as well (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 which will ensure reliable coverage on most days (as we did for sexual desire, a central construct in our study). Overall, we are confident that these planning costs are worth paying, if they allow us to answer questions about measurement reactivity, a central concern that affects much of psychological measurement.

A good rule of thumb may be that any measurement-related issue that comes up as a topic on which team members disagree when planning a study is a candidate for randomisation, so that disagreements can be resolved by data. Teams may have to get over a certain degree of experimentation aversion (Meyer et al., 2019) to do so, but should remind themselves that if they worry about a design question affecting their results, preferring ignorance of the consequences is not a reasonable strategy. Candidates for randomisation could additionally include item wording, the order and the number of response categories for multiple choice questions. In addition to the analyses we conducted here, future work might then a) additionally estimate potential elevation biases at the survey level by randomising start dates after

recruitment to find out whether starting a new survey causes initial elevation bias, b) investigate the impact of (randomised) measurement frequency and participant burden on dropout, nonresponse rates, and data quality (Little & Rhemtulla, 2013), c) conduct cognitive modelling to bring together the changes in response times and answers in a coherent framework, and d) estimate empirically and theoretically grounded exclusion thresholds for overly fast responses, where the thresholds are dependent on normative curves for the amount of experience participants have had with a survey and item (it seems that current practice tends to use fixed thresholds).

In summary, to avoid flying blindly with respect to the potentially deleterious effects of measurement reactivity, researchers should routinely randomise measurement frequency and order at the item and survey levels when using subjective reports, whether they are repeated or singular, clinical or population-level. Researchers who already use designs with planned missingness without randomisation should consider adding randomisation to not just reduce participant burden but to also learn about its effects (Little & Rhemtulla, 2013). Doing so can lead us from a culture where we estimate measurement reactivity correlationally, discuss it in footnotes, and hope for the best to a culture where we randomise, estimate, and adjust for measurement reactivity to reduce and prevent these biases.

Author contributions

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

Acknowledgements

We thank the participants of our diary study and the Leibniz ScienceCampus Göttingen which partially funded the study.

References

Arslan, R. C. (2018). *Initial elevation bias analyses https://rubenarslan.github.io/initial_elevation_bias/*. https://doi.org/10.5281/zenodo.1254127

Arslan, R. C., Walther, M., & Tata, C. (in press). formr: A study framework allowing for automated feedback generation and complex longitudinal experience sampling studies using R. *Behavior Research Methods*. https://doi.org/10.31234/osf.io/pjasu

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research*. Retrieved from https://market.android.com/details?id=book--sf2wP9-T98C

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

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

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. Retrieved from 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

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*. Retrieved from 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., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavél, F. D., … 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