Is it Possible to Identify Careless Responses with Post-hoc Analysis in EMA Studies?

Jana Welling
WS Audiology
jana.welling@gmail.com

Rosa-Linde Fischer
WS Audiology
rosa-linde.fischer@wsa.com

Nadja Schinkel-Bielefeld
WS Audiology
nadja.schinkel-bielefeld@wsa.com

ABSTRACT

Data quality is a major issue when conducting studies in behavioral sciences. One of the possible threats to data quality in user modeling, in particular in questionnaire studies, is providing careless responses (CR). When responding carelessly, subjects do not pay sufficient attention to the questions and therefore compromise the interpretability of the responses. The aim of the current study was to gain a better understanding of the occurrence and identification of CR in Ecological Momentary Assessment (EMA) studies, where several questionnaires usually are administered daily to the participants over the course of some days, weeks or even months. For this purpose, explorative post-hoc analysis was conducted using the data of an existing EMA study in audiological research. Completion time, variance, skipped items, acquiescence bias and number of textboxes were analyzed as potential indicators for CR both inter- and intraindividually. Furthermore, consistency was examined using linear mixed models and scanning individual questionnaires. Results showed minimal systematic inconsistencies, indicating the absence of large-scale CR. However, this type of analysis might not be appropriate for identifying CR when only occurring occasionally. Moreover, the reliability of indicators of CR might be limited in EMA studies, as the indicators also vary over the course of the study and between different situations. Possibilities for future studies are discussed.

CCS CONCEPTS

• **Human-centered computing** → Human computer interaction (HCI); HCI design and evaluation methods; User studies; Human computer interaction (HCI); Empirical studies in HCI; Ubiquitous and mobile computing; Empirical studies in ubiquitous and mobile computing.

KEYWORDS

data quality, careless responses, Ecological Momentary Assessment, Experience Sampling, acquiescence bias

ACM Reference Format:
Jana Welling, Rosa-Linde Fischer, and Nadja Schinkel-Bielefeld. 2021. Is it Possible to Identify Careless Responses with Post-hoc Analysis in EMA Studies?. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21 Adjunct), June 21–25, 2021, Utrecht, Netherlands.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3450614.3462237

1 INTRODUCTION

Personalization and user modeling require user input, either in the form of observed behavior or in the form of questionnaire responses. To gain valuable information from this user input, it is essential to ensure the quality of the collected data. The current study focuses on the data quality of questionnaire responses in Ecological Momentary Assessment (EMA) studies. In these studies subjects usually are asked to complete several questionnaires per day over the course of several weeks often accompanied by passive sensing (e.g. [1, 2]). However, responding to questions in surveys involves different cognitive processes that demand effort from the respondents [3]. If respondents are not sufficiently motivated to provide optimal answers, they may respond without due diligence and therefore give careless responses [4]. This can lead to lower quality data and hence complicate drawing conclusions or modelling relevant behavior based on such data [5]. Although several studies have investigated the occurrence and identification of careless responding (CR) in cross-sectional research designs [4, 6], research on CR in the area of EMA studies is scarce [5]. With the popularity of the EMA method in behavioral sciences, further investigations into the occurrence of CR in these studies should provide valuable information for future research. Therefore, the aim of this study is to expand on the topic of CR in EMA studies. In the following sections we will discuss existing research and its generalizability to EMA studies.

1.1 What is CR and how often does it occur?

Careless responding is defined as responding without paying sufficient attention to the questions [4]. Its form can vary from extreme response patterns such as straightlining (always choosing the same answers) to completely random answers [4, 6, 7]. Research has shown that there is a moderate to high prevalence (around 5 – 20%) of CR in cross-sectional studies [4, 6, 8]. However, to our knowledge only one study exists that investigates CR in EMA. Eisele and colleagues [5] triggered their participants up to nine times per day over the course of two weeks to respond to a questionnaire which could be either short (30 items) or long (60 items). They included subjective and objective measures of CR in their questionnaires. Participants indicated to have responded mostly with care, as the median response was 6 out of 7 on a 7 point Likert scale (with 7 indicating full attention). Similarly, only few participants (3.6%) failed to answer in an attention test with "not at all" when specifically instructed to do so. Careless responding was not affected by length of questionnaire, frequency of triggers or time over the course of the study. Generally, protecting factors against CR are high age

UMAP '21 Adjunct, June 21–25, 2021, Utrecht, Netherlands
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8367-7/21/06.
https://doi.org/10.1145/3450614.3462237

and self-esteem, agreeableness, conscientiousness and openness, as well as an interest in the research topic and intrinsic motivation [6].

1.2 How can CR be identified?

There are two ways of identifying CR. First, researchers can include a priori items or entire scales in the study design to detect the care or carelessness with which participants are responding to the questions. One approach is to include items that instruct the participants to give a particular response, with the instructions usually "hidden" in the question text. Such items are quite effective in detecting CR in cross-sectional studies [6, 7]. However, subjects in EMA studies answer several questionnaires daily and might therefore recognize such items over time and possibly be annoyed or insulted by the apparent distrust in their responding behavior. Another approach is to ask subjects directly – either in the current questionnaire or retrospectively – whether they engaged in CR. Although this technique seems quite straightforward, it could be affected by subjective bias. In particular, the item itself might be subject to CR and therefore not reliable. Furthermore, in EMA studies, it is especially important to keep questionnaires short to minimize the burden for the participants. Thus, including extra items might not be advisable. As a second option, data can be analyzed post-hoc to search for conspicuous context parameters or response combinations [7, 8]. Statistical values like completion time, ratio of missing items, half-split reliability, Mahalanobis distance or variance can be calculated and used to identify CR [7, 8]. However, most of these values are only able to detect certain kinds of CR (e.g. straightlining). Consequently, combining these values to an index for CR identification has limited effectiveness [6]. To compare the effectiveness of different statistical values in distinguishing between careless and careful responses, Leiner [7] conducted an experiment based on a cross-sectional research design. While one of two groups was instructed to respond to survey questions carefully, the other group was instructed to respond carelessly. Only the completion time was effective in distinguishing between both groups. Variance was only effective in detecting straightlining.

To our knowledge there exists no study about post-hoc identification of CR in EMA studies yet. However, we assume that the unique structure of the data offers possibilities and poses challenges to the post-hoc analysis. On the one hand, the repeated measures design allows the analysis not only on an interindividual level, but also on an intraindividual level. Thus, it might be possible to search both for persons that tend to generally engage in CR and for single questionnaires with an increased amount of CR. On the other hand, the design might render some of the statistical values less reliable in identifying CR. For example, the time needed to complete a questionnaire could depend on the time elapsed since the start of the study, as subjects become familiar with the questions, and consequently be less indicative of CR. Also, divided attention between responding to the questionnaire trigger and continuing engagement in the current activity might lead to longer response times in some situations. Last, because of the large number of questionnaires, it is easier to skip a questionnaire than in cross-sectional studies. Thus, subjects may rather omit responding than responding carelessly, which still requires some effort.

A main goal of the current study is to get general insights into the possibilities of identifying CR in EMA studies by post-hoc analysis. Second, we wanted to be able to draw a more informed conclusion about the occurrence of CR in our kind of studies. The respective analyses are based primarily on descriptive and explorative methods. Special characteristics of the methods used in the original study and its consequences for the analysis are being discussed.

2 METHODS

2.1 Participants and procedure

The analysis is based on data of an EMA study in which 20 participants with moderate hearing loss were surveyed over the course of three weeks (for more details see [9]). Participants were 24 to 82 years old (M = 67.5, SD = 17.0) and mostly recurrent study participants. They were provided with smartphones that included a previously installed EMA app. During the home trial, participants completed on average 182 questionnaires (SD = 72) via the app. Three types of triggers existed to start a questionnaire. First, participants were randomly triggered to start a questionnaire eight times per day. Second, if they were in a loud environment (> 65 dB SPL), they were triggered up to four more times per day. Third, participants could always initiate questionnaires themselves. Data on the acoustic situation (e.g. level, classified listening situation) was collected continuously by the hearing device. Subjects received reimbursement for their travel to the laboratory and were paid 12 € per hour for the initial and final session. They received 20 Cents for each completed short questionnaire (up to 7 questions) and 70 Cents for each completed long questionnaire, with a maximum of 10 € per day. No subject completed enough questionnaires to reach maximum daily reimbursement.

Each questionnaire consisted of a mix of single choice (SC), multiple choice (MC), slider (SL) and textbox (TB) questions with varying number of response options (2-8) for each question. No scale consisting of more than one item was used. The number of questions changed according to the subject's responses. In particular, subjects could end the questionnaire after five to eight mandatory questions. The remaining questions were optional (e.g. single questions could be omitted). The length therefore varied between five and 31 questions (M = 14.7). Subjects were able to pause the questionnaire for up to 30 minutes. After that time the remaining questionnaire was dismissed and identified as "timeout". A list of the questions used in the questionnaire can be seen in Schinkel-Bielefeld et al [7].

2.2 Post-hoc analysis

For the analysis, five potential indicators of CR were identified. As Leiner [7] found the completion time to be the most effective indicator of CR, this parameter was included in the analysis. However, as subjects were instructed, to open the questionnaire and return to it later if the situation was inopportune, and because the length of the questionnaire and type of questions varied, total completion time is presumably no very reliable measure. To account for this, questionnaires with more than 10 minutes of total completion time were deleted from the time analyses. Furthermore, the median was used for analyses instead of the mean whenever possible. Completion time will hereinafter refer to the average completion time per question.

As a second potential indicator of CR, average variance per questionnaire was identified, although in the study by Leiner [7] variance was reported as only effective in detecting straightlining. To create this parameter, only questions with response scales similar to Likert scales were used. To account for the different scale sizes, each scale was adjusted post-hoc to create equal scales with a (hypothetical) mean of 0, a (potential) maximum of 1 and a (potential) minimum of -1. E.g., if a scale consisted of six answers, answer six would be represented by 1, answer four by 0.2 and answer two by -0.6. This adjustment was independent of the actual answers provided by the subjects. In the end, the variance within each questionnaire was computed.

The average number of skipped items (as suggested in the study of Barge and Gehlbach [8]) and the average number of textboxes within a questionnaire were also identified as potential indicators of CR. TB questions only appeared after a certain answer to the previous question. Therefore, after familiarization with the questionnaire, subjects might start to avoid answers that trigger TB questions. and hence render the average number of textboxes an indicator of CR.

Another consideration for survey questionnaires is the nature of participants to choose “agree” to most agree/disagree questions in a survey, even when questions might contradict each other. This bias is called acquiescence bias [10] and was treated as a last potential indicator of CR. Although this bias itself can be a problem to data quality, it might help in identifying CR. If an acquiescence bias exists, and if its size is equal in normal scaled and inversed items, this would suggest that subjects read the questions and response options with care and answer accordingly.

Post-hoc analysis was of an explorative nature and based mostly on descriptive analysis. Completion time, variance, skipped items, acquiescence bias and number of TB questions were analyzed both inter- and intraindividually. Provided answers were also compared with the acoustic data. Furthermore, consistency analysis was conducted using linear mixed models (LMM). In addition, scatter plots of responses to similar questions were created to reveal inconsistent response combinations, e.g. poor speech understanding and low listening effort. These questionnaires were then scanned individually for further inconsistent content. Analysis was conducted using Matlab R2016b, whereas LMMs were conducted in R version 4.0.3 with the *nlme* package [11].

3 RESULTS

3.1 Descriptive findings

Participants stayed in the home trial between seven and 33 days (M = 21.0), completing on average between 2.5 and 12.8 questionnaires per day (M = 8.8). On average, 4.6 of these were random-triggered. As each subject received eight random triggers per day, the average response rate was 58 %. Participants further filled out 2.9 self-triggered questionnaires and 1.3 questionnaires triggered from loud environments per day. Subjects ended 2 to 40 % of the questionnaires after the mandatory questions (M = 13 %). Overall, participants skipped between 0 and 0.42 questions per questionnaire (M = 0.06). When excluding all questionnaires with a total completion time of more than ten minutes, the median completion time per question lay between 2.7 and 19.2 seconds (M = 7.3), therefore varying strongly between participants. The answers to SC questions varied on average between 0.10 and 0.59 points of the standardized scale (M = 0.33). The questionnaires contained on average 0.24 TB questions, with subjects responding to between 93 and 100 % of them.

3.2 Do people generally engage in CR?

First of all, we investigated whether some of the participants engaged in CR behavior throughout the EMA questionnaires. For this purpose, we analyzed potential indicators for CR on the person level. Figure 1 shows a scatter plot indicating median completion time, average variance, average number of skipped items and number of self-triggered questionnaires per subject throughout the study. We expected subjects who engaged continuously in CR to show a small median completion time, probably a small average variance (note, however, that variance was only able to effectively detect straightlining in the study by Leiner [7]), a high number of skipped items and – if they were participating in the study solely for the reimbursement – a high number of self-triggered questionnaires. As can be seen in Figure 1, no subject shows such a combination of all four indicators. However, subject 12 combines a low completion time and variance with a high number of self-triggered questionnaires. Subjects 4, 11 and 17 have a relatively high number of skipped questions per questionnaire, whereas subjects 1 and 9 responded to the questionnaires quite fast.

If subjects responded randomly to the questions, there should be no bias towards agreement to the questions. Therefore, we treated the absence of an acquiescence bias as an indicator of CR. The tendency of the subjects' responses can be compared between normal scaled items (confirming responses at the top of the screen) and inversed items (negating responses at the top of the screen). If subjects were engaging in CR, they should either show no acquiescence bias at all or tend to answer all questions in the same direction, no matter what the content of the question or response options is. Therefore, we should expect a difference in acquiescence bias between these two groups of items. Figure 2 shows the acquiescence bias of all subjects for both normal scaled and inversed items. Accordingly, each subject generally shows an acquiescence bias in his or her responses, indicating the absence of completely random CR. Furthermore, most subjects differed in their bias between the two groups of items, indicating the presence of not random CR. Bootstrap analysis confirmed this result statistically (for all except subject 15: $p < .05$). However, note that the two groups of items differed also in their content – the normal scaled items were questions about the personal state of the participant (e.g., his or her mood), whereas the inversed items referred to the acoustical situation and the satisfaction with the hearing aids. Therefore, the difference in acquiescence bias might rather be attributed to the distinct content. This assumption is supported by the results: all subjects except three (5, 13 and 15) rather affirm to questions regarding their personal state than to questions regarding the acoustical situation. If the results were to indicate CR, the direction of the difference should be equally distributed over the two groups of items.

3.3 Do people sometimes engage in CR?

Subjects might not show a general tendency of engaging in CR, but rather do so only from time to time or in certain occasions. For this purpose, we not only used inter- but also intraindividual analysis

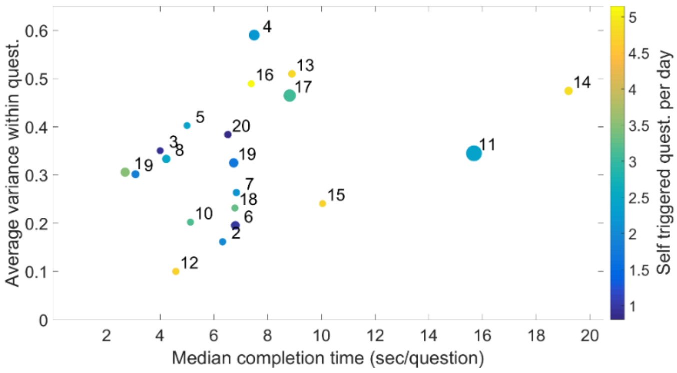

Figure 1: Average variance, median completion time, average number of self-triggered questionnaires per day and average number of skipped items per questionnaire (0-0.42, indicated by the size of points). Displayed numbers denote the subject.

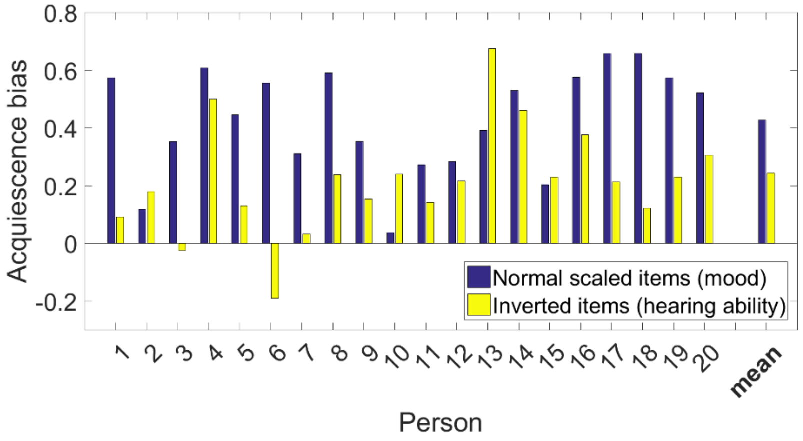

Figure 2: Acquiescence bias for normal scaled and inverted items.

to identify potential CR in the data. First, histograms were created for each subject showing either the completion time, variance or skipped items over all questionnaires. If subjects engaged only sometimes in CR, the intraindividual distribution of these indicators could show some extreme outliers or even be bimodal. The graphical analysis illustrated that for all subjects there were no indicators of a clear bimodal distribution, indicating the absence of large-scale CR. Furthermore, scatter plots were created plotting the completion time against variance for each subject. If subjects engaged only occasionally in straightlining as indicated by questionnaires with both low variance and completion time, we should expect to see a positive relationship between these two variables and an accumulation of questionnaires with low variance and completion time. The graphical analysis yielded no such results, indicating the absence of CR.

CR in EMA studies might depend on the progress of time in the study and the situations in which questionnaires are filled out. Subjects might become exhausted with time or be less motivated in certain situations. In this case we should expect to see different mean values in the indicators of CR in different weeks or situations. Therefore, the median completion time, the average variance, the number of skipped items and the textboxes were plotted over the three weeks of the study and different encountered situations. Subjects 3, 6, 14, 15 and 19 were excluded from analysis due to early termination of the study (less than three weeks). The graphical analysis was supported by three-way ANOVAs with time, situation and person as factors and the respective indicator as the dependent variable. Results demonstrate that only completion time decreased over the course of the study. As no other indicator changed over time, this might be due to a training effect rather than to the engagement in CR. However, all potential indicators of CR differed between situations. Figure 3 shows the median completion time over the course of the study and in different situations. As can be seen, completion time seems to be higher in demanding conversations, e.g. on the phone or in group conversations, than in other situations. For the other indicators no clear trend is identifiable.

Furthermore, subjects might tend to engage in CR either when they initialize the questionnaire themselves (e.g., they fill out questionnaires only for reimbursement) or when the app initializes the questionnaire (e.g., the questionnaire is started in an inopportune

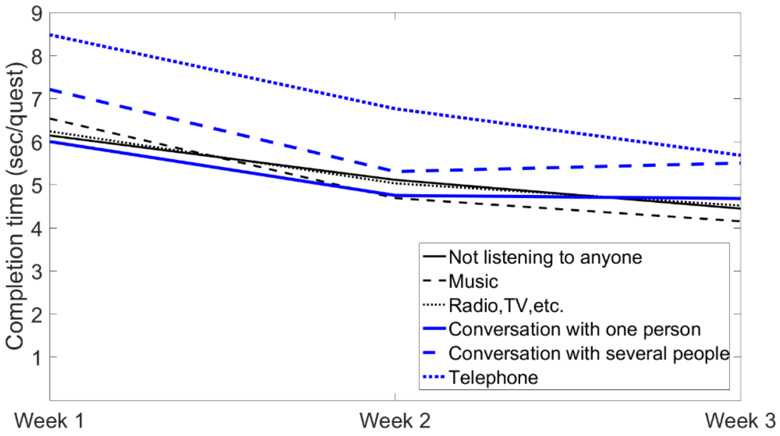

Figure 3: Completion time in dependence of week of study and listening object.

situation). Therefore, we compared the different indicators for CR between the three types of triggers to detect potential differences. Two-way ANOVAs were conducted with trigger and person as factors and the respective indicator as dependent variable. Only with completion time as dependent variable a main effect for trigger could be found ($F(2) = 31.3$, $p < .001$). The graphical analysis indicated a slightly higher completion time for self-triggered questionnaires, suggesting more carelessness when filling out randomly triggered questionnaires.

3.4 Consistency analysis

The questionnaires of the study included some questions with similar content, e.g. regarding satisfaction with the hearing aid in general and with sound quality and speech comprehension. Responses to these questions are expected to show a positive relationship, as poor sound quality or poor speech intelligibility may reduce the overall satisfaction with the hearing aid. Accordingly, overall satisfaction was significantly and positively correlated with both sound quality (r = .69, p < .001) and speech comprehension (r = .49, p < .001). Similarly, the background noise indicated by the subjects was significantly and positively correlated with the mean level during the first three minutes of the questionnaire assessed by the hearing aid (r = .33, p < .001). These 'natural' relationships can be used to statistically test whether consistency in questionnaires varies depending on different indicators of CR. Whenever subjects don't pay sufficient attention to the questions, these relationships should be weaker than usual. Therefore, if the subjects indeed engaged in CR, it is expected that the CR indicators moderate the aforementioned relationships.

To test this assumption, two linear mixed models (LMM) were conducted. In the first LMM, overall satisfaction with the hearing aid was predicted by speech comprehension and sound quality. In the second LMM, the (subjective) background noise was predicted by the (objective) mean level. In both models, random intercepts and random slopes for the original predictors were introduced with subjects as a grouping variable. As a last step, the following indicators of CR and their interaction terms with the original predictors were added to the model: completion time, variance and a dummy variable indicating whether the subject triggered the questionnaire him- or herself (1: self-triggered, 0: rest). Number of skipped items and of textboxes were not included in the analysis as these variables were strongly right skewed with an average close to zero.

In the first LMM, speech comprehension (Est. = -0.20, SE = 0.04, $t = 4.64$, $p < .001$), sound quality (Est. = 0.33, SE = 0.06, $t = 5.28$, $p < .001$) and the interaction of speech comprehension and variance (Est. = 0.10, SE = 0.05, $t = 1.97$, $p = .049$) significantly predicted overall satisfaction with the hearing aid. High speech comprehension and sound quality therefore go along with high overall satisfaction with the hearing aid. Furthermore, speech comprehension had a stronger statistical effect on overall satisfaction when variance in the questionnaire was high, indicating the possibility of systematic CR. In the second LMM, mean level (Est. = 0.03, SE = 0.01, $t = 5.00$, $p < .001$) and variance (Est. = -2.84, SE = 0.66, $t = -4.30$, $p < .001$) predicted the background noise indicated by the subjects. High mean level and low variance in the questionnaire therefore go along with an indication of loud background noise. In both models, no other variables or interaction terms were significant predictors of the outcomes. Moderation effects by the indicators of CR could therefore mostly not be confirmed.

In addition to the statistical consistency analysis, scatter plots were used to search for inconsistent answers in pairs of similar questions. Comparisons were made between speech comprehension and listening effort, sound quality and overall satisfaction, as well as two questions regarding the surroundings. The respective questionnaires were then individually scanned for further inconsistencies or explanations of inconsistencies. This method yielded few inconsistent response combinations, e.g. "I am currently walking/biking" and "I am currently using public transportation". However, most of these inconsistencies could be explained when scanning the questionnaire, e.g. because the situation had changed while filling out the questionnaire. Moreover, subjects occasionally filled out lengthy textboxes to explain the current situation, rendering carelessness unlikely. Only subject 11 exhibited few not easily explainable inconsistencies in a total of 138 questionnaires, such as indicating silence in the background, but stating that the background noise is loud. For subjects 1, 4, 9, 12 and 17, who had high values in various indicators, no unexplainable inconsistencies were noted in their questionnaires.

4 DISCUSSION

We had two main goals in the current study. First, we wanted to get more insight into the possibilities of identifying CR in EMA studies using post-hoc analysis. Second, we wanted to be able to draw a more informed conclusion about whether CR exists in the kind of EMA studies we conduct. The analyses showed barely any hints of CR in the study data. Interindividually, only one subject (11) exhibited both conspicuous values in some of the indicators of CR and inconsistencies in content. However, this subject was exceptionally slow in completing the questionnaires (Figure 1). As this person was with 79 years one of the oldest subjects, at least some of the conspicuous results might be caused by general difficulties with the smartphone or other aspects of the study procedure.

Acquiescence bias existed in the data and was similar – but not equal – between normal scaled and inverted items, indicating the absence of systematic CR. Intraindividually, no bimodal distributions were found, indicating that none of the subjects engaged in CR on a large-scale basis. Furthermore, motivation does not seem to decrease over time, as indicators of CR – except for completion time – did not change over the course of the study. However, indicators did differ between situations. As these differences were not consistent across indicators, it is unlikely that they depict the tendency of persons to engage in CR rather in one situation than the other. In fact, this result might – as well as the decreasing completion time – rather suggest that the sole reliance on statistical parameters for post-hoc identification of CR in EMA studies could be problematic. In contrast to cross-sectional studies, in EMA the parameters seem to vary naturally between situations and (in the case of completion time) over the course of the study. Differences in the potential indicators of CR might therefore reflect different situations and time points rather than questionnaires with and without CR. Thus, the indicators identified in studies based on cross-sectional data might not be reliable and valid indicators in EMA studies.

Subjects needed slightly more time to complete self-triggered questionnaires than for random-triggered questionnaires. If completion time was a reliable and valid indicator of CR, this might suggest that CR occurs more prevalently in random-triggered questionnaires. This seems reasonable, as subjects could be presented with random triggers in inconvenient situations, causing them to rush through the questionnaire. However, as we cannot be sure about the reliability and validity of completion time as an indicator of CR, we should only carefully interpret the results. Subjects were instructed to fill out questionnaires in many different situations. Thus, it is also plausible that more time is required to complete self-triggered questionnaires as they are more often completed in rare and difficult situations.

Statistical consistency analysis did reveal a moderating effect of variance on the positive relationship between satisfaction and speech comprehension. As this was the only moderation effect found, we should be careful with our interpretation regarding CR. Note, however, that a non-significant result is no evidence for the nonexistence of the effect, as it highly depends on the power of the analysis. The non-significant results should be therefore interpreted only as an indication towards the absence of CR in the screened study. Individual consistency analysis revealed barely any inconsistencies in the questionnaires. As discussed above, only subject 11 showed few inconsistencies which could not be explained.

As the data shows only few systematic or individual inconsistencies, presumably there is no large-scale or systematic CR in the screened study. This conclusion corresponds with several theoretical reasons not to expect many CR in EMA in general or in the screened study in particular. Questionnaires were relatively short compared to surveys used in cross-sectional studies, with the possibility to not answer at all, to stop the questionnaire after a few mandatory questions or to pause the questionnaire and return to it later. Furthermore, as subjects in audiological research studies usually have a hearing loss, they are more likely to be intrinsically motivated to participate in the study and have a genuine interest in the research topic, both protecting factors against CR [6]. Additionally, the participants in this particular study were experienced study participants and consequently familiar with the investigators. Careless responses might therefore be no big problem in EMA studies similar to ours. Selection bias and missing responses might be far greater threats to data quality, as subjects can choose in which situations to answer to the questionnaires and in which not.

The design of this study does not allow any certain conclusions. The methods used in this study are best suited to detecting systematic and large-scale CR, if they are reliable and valid in EMA studies at all. Furthermore, most indicators and analyses might only be able to detect certain types of CR, rendering the identification of CR based solely on post-hoc analysis difficult. Another important drawback is the general limited knowledge about CR in EMA studies to date. As we do not know how many and what kind of CR to expect, we actually do not know what to look for. E.g., do all persons sometimes engage in CR or does something like a "CR personality" exist where individuals tend to engage more regularly in CR than others?

To answer these open questions, future studies are needed that investigate CR in EMA studies more thoroughly. In quasi-experiments, subjects could be randomly instructed at the beginning of each questionnaire whether to respond to the questions with care or without paying attention to them. Similar to the Leiner study [7], potential indicators of CR could be reviewed and reliable (post-hoc) methods of detecting CR identified. These methods could be used in future studies to answer the open questions regarding CR in EMA studies. Another possibility for future studies, similar to the study of Eisele and colleagues [5], would be to ask subjects either immediately at the end of each questionnaire or retrospectively at the end of the study whether they actually engaged in CR or not. Note, however, that this technique might upset subjects, as they could be disappointed by the distrust in their compliance. Furthermore, questionnaires might be adjusted in future studies to simplify post-hoc analysis. E.g., questions could be adapted to have the same or at least same number of response options and completion time could be measured per question. Moreover, some of the items could be randomly inversed, making it easier to detect straightlining.

5 CONCLUSION

Despite the limitations of the current study, it contributes to the better understanding of CR in EMA studies with its wide array of analyses including both quantitative and qualitative approaches. We are convinced that CR is an important topic which should attract more attention in EMA research. Even though CR might not be problematic in small studies with personally known subjects, it could be more relevant for large scale studies that recruit subjects without personal contact, e.g. via an external service provider or when subjects can participate by simply downloading an app on their smartphone. Therefore, the identification and avoidance of CR are a major opportunity to improve data quality and draw more valid and reliable conclusions.

ACKNOWLEDGMENTS

We thank Dr. Eric Branda for carefully proofreading the manuscript.

REFERENCES

[1] Rhin, S., Lee, U., & Han, K. (2020). Tracking and Modeling Subjective Well-Being Using Smartphone-Based Digital Phenotype. In Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (pp. 211 – 220).

[2] Doherty, S. T., Lemieux, C. J., & Canally, C. (2014). Tracking human activity and well-being in natural environments using wearable sensors and experience sampling. *Social Science & Medicine*, *106*, 83-92.

[3] Krosnick, J. A., & Presser, S. (2010). Question and Questionnaire Design. In P. V. Marsden & J. D. Wright (Eds.), Handbook of survey research (pp. 263-313). Emerald Group Publishing.

[4] Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. Psychological methods, 17(3), 437.

[5] 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.

[6] Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. Journal of Research in Personality, 48, 61-83.

[7] Leiner, D. J. (2019). Too Fast, too Straight, too Weird: Non-Reactive Indicators for Meaningless Data in Internet Surveys. In Survey Research Methods (Vol. 13, No. 3, pp. 229-248).

[8] Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. Research in Higher Education, 53(2), 182-200.

[9] Schinkel-Bielefeld, N., Kunz, P., Zutz, A., & Buder, B. (2020). Evaluation of hearing aids in everyday life using Ecological Momentary Assessment: What situations are we missing?. American Journal of Audiology, 29(3S), 591-609.

[10] O'Muircheartaigh, C. A., Krosnick, J. A., & Helic, A. (2001). Middle alternatives, acquiescence, and the quality of questionnaire data. Chicago, USA: Irving B. Harris Graduate School of Public Policy Studies, University of Chicago.

[11] Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2020). _nlme: Linear and Nonlinear Mixed Effects Models_. R package version 3.1-151, <URL: https: //CRAN.R-project.org/package=nlme..