An Attempt to Detect Social Desirability Bias in AI Chatbot Interaction and Privacy

D. Farinella & M.M. Callejo

April 13, 2025

Abstract

This study originates from Weisgarber et al.' experimental research on the influence of social desirability (SD) bias on self-report online surveys (Weisgarber, Valacich, Jenkins, Kim, & Kumar, 2024). The authors employed Human-Computer Interaction (HCI) dynamics such as mouse-tracking as a tool to get insights on the respondents' behavior. In the present study we replicate the original experiment and address one main limitations acknowledged by the authors themselves. Participants (n=39) filled out a questionnaire on AI chatbot interaction and privacy while their mouse movements were tracked. Our findings indicate that, while in the original on health and wellness study participants who were prompted to be more affected by SD bias showed higher response times and slower mouse cursor speed, in the context of AI chatbot interaction and privacy HCI dynamics fail to provide valuable insights on the participant's SD bias. All in all, these results suggest that the method proposed by Weisgarber et al. needs to be based on more solid experimental results and is not generalizable.

Keywords: Social desirability bias, HCI dynamics, mouse-tracking, AI, chatbots, cybersecurity, online survey research

1 Introduction

As Weisgarber et al. acknowledge, collecting self-reported data is undoubtedly a fruitful and widely used method in research and experimental literature. In particular, this self-reported answers (collected by means of surveys) not need be directly observed by the researcher as observational data. This simple and basic feature of self-reported answers constitutes a huge advantage in the context of experimental research, due to two core reasons. As for the former, being observational data directly collected, the risk of the researcher not fully embracing the phenomenon occurs. The latter lies in the fact that observational data cannot be collected from huge populations (due to time and resources limitations) and hence contitutes a non-scalable method. However, Weisgarber et al. argue that, despite representing an efficient and scalable method, surveys do suffer from two important limitations.

Initially, they are affected by the impact of social desirability (SD) bias. In short, SD bias prompts respondents to overproduce social desirable responses and underproduce social undesirable ones (Bhattacherjee, 2012).[1] The impact of SD bias has been investigate by experimental literature in many areas (e.g., health, political ideas, cybersecurity, and so on).[2]

Furthermore, surveys do not provide any insights into the respondents' behavior during the decision-making process. That is, researchers are able to log the final decision but have no clue about how the participant acted *while* filling the survey. This limitation is even sharper in the case of online questionnaires. As literature suggests (Weinmann, Valacich, Schneider, Jenkins, & Hibbeln, 2022), tracking subjects behavior during online tasks by means of Human Computer Interaction (HCI) dynamics can dis-

[1] The author includes the SD bias among the five systematic that can invalidate experimental studies and describes it as follows: ≪Many respondents tend to avoid negative opinions or embarrassing comments about themselves, their employers, family, or friends. [...] This tendency among respondents to "spin the truth" in order to portray themselves in a socially desirable manner is called the "social desirability bias", which hurts the validity of response obtained from survey research≫.

[2] See e.g., (Hoffman, Aden, Barbera, & Tvaryanas, 2023), (Tappin, van der Leer, & McKay, 2023)

close relevant and conceived perspectives.

Weisgarber et al. investigated the ability of HCI dynamics (such as mouse movements metrics) to provide insights on the respondents' behavior during a self-reported survey on health and wellness. In their study, participants were asked to complete a four sections health and wellness questionnaire while their mouse movements were tracked. Importantly, the social desirability bias of the respondents was manipulated by different instructions. The researchers found that mouse movements metrics can provide significant insights into the effect of SD bias on the participant's response, and concluded that their approach constitutes a scalable method for identifying and addressing SD bias effect in online surveys. However, as Weisgarber et al. themselves acknowledge, their experimental study does suffer from several limitations. Namely, (1) it employs respondents from a specific demographic; (2) it uses questions related to a specific topic (health and wellness); (3) it does not check for other variables that could affect the HCI dynamics measured; (4) being a between-subject design, it does not take the variability across subjects into account.

In this study, our aim is to address the second (2) limitation of Weisgarber et al.' experimental analysis by investigating if HCI dynamics such as mouse tracking can provide insights on the effect of SD bias also in a different field: AI chatbot interaction and privacy. Basically, we replicated Weisgarber et al.' and asked partecipants to complete a two sections AI chatbot interaction and privacy online-survey while their mouse movements were tracked. Again, the respondents' SD bias was manipulated by different instructions. Slightly modifying the original study's one, we address the following research question: *Do respondents' mouse movements differ in AI chatbot interaction and privacy surveys when social desirability is manipulated through two experimental conditions?*

2 Theoretical and Experimental Background

In this section, we briefly present the necessary rudiments as regards the identification and control of SD bias in experimental study as well as the line of thought followed by Weisgarber et al. in the development of their research question. Then, we summarize their original experiment and its findings.

2.1 Social Desirability Bias: Scales, Mitigation and Identification

Being SD bias at the center of this experimental study, we find it necessary to provide some theoretical background about it. In order to do so, this section serves as a brief introduction to three main topics related to SD bias: (1) the scales that aim at measuring it; (2) the techniques developed to minimize it; (3) the possibility to track it by means of HCI metrics.

The possibility of measuring the tendency of subjects to provide SD responses has been investigated from decades by academic literature. This line of research mostly consists in building *scales* to measure the SD tendency of subjects. Among the multitude of alternatives,[3] Weisgarber et al. present the scale developed by Crowne & Marlowe (Crowne & Marlowe, 1960) as the most reliable one and hence employ it in their experiment. So as to truthfully replicate their method, we do the same. Crowne & Marlowe' approach develops a scale by posing the subject questions about their behavior in everyday situations.[4]

As regards mitigation techniques for SD bias, Weisgarber et al. highlight two main approaches. The former consists in giving the respondent the perception of anonimity (wheather this is real or not). Apparently, if participants feel like their answers are going to be kept anonymous, SD responses are less likely to be produced. As for the latter approach, experimental literature has shown that instructions such as 'we evaluate your answer' or emphasis on honest and pondered responses can considerably mitigate the impact of SD bias on participants. Of course, the efficiency of such techniques can vary greatly depending on the context.

[3] See e.g., (Schuessler, Hittle, & Cardascia, 1978), (Paulhus, 1988), (Steenkamp, De Jong, & Baumgartner, 2010).

[4] Some of the items are, e.g., "I like to gossip at times", or "I am always courteous, even to people who are disagreeable.". The answers are provided as a truth or false choice.

Finally, the last theme addressed in this paragraph regards the possibility to track SD bias through HCI dynamics and in particular through mouse-tracking metrics. Weisgarber et al. acknowledge that, although mouse-tracking metrics habe been employed to analyze different types of phenomena in experimental literature,[5] SD bias was not among those. Ultimately, the lack of HCI analysis on SD bias motivated the authors to investigate this specific topic. Decided in attempting to track social desirability bias through mouse cursor movements, the authors elected health and wellness as a suitable domain to run their analysis.[6]

2.2 Cognitive Control as a Core Process

The possibility to track SD bias through mouse cursor movements, from a theoretical standpoint, is connected with a line of research that conceives *cognitive control* as an important feature in modulating humans' behavior. In short, this line of though suggests that individuals are intially inclined to respond truthfully. Only after, when cognitive control activates, they experience the need to hide socially undesirable behaviors, and hence to lie.

Now, HCI dynamics seems to provide insights on respondents' cognitive control (and thus on their SD bias). In particular, Jenkins et al. showed that respondents tend to move slower and perform more deviations when trying to hide information during an answering process that involves the use of a mouse cursor (Jenkins, Proudfoot, Valacich, Grimes, & Nunamaker, 2019).

2.3 SD bias measures through mouse cursor movements

The original experiment of Weisgarber et al. based on the line of research suggested above originates from a clear conviction: if respondents are prompted to be more effected by SD bias (e.g., through non-anonymity and omitting requests for honesty) they are likely going to be less confident and reflect more; quite the opposite, if respondents are prompted to be less effected by SD bias (e.g., through anonymity and requests for honesty) they are likely going to be more confident and reflect less. Based on this conviction, the authors construct the four research hypotheses reported below:

H1. *Respondents in the higher social desirability group will exhibit longer response times compared to the respondents in the lower social desirability group.*

H2. *Respondents in the higher social desirability group will exhibit greater mouse cursor deviations compared to the respondents in the lower social desirability group.*

H3. *Respondents in the higher social desirability group will exhibit slower mouse cursor speeds compared to the respondents in the lower social desirability group.*

H4. *Respondents in the higher social desirability group will exhibit more answer switches compared to the respondents in the lower social desirability group.*

2.4 A mouse tracking study on SD bias

Having clarified the theoretical background as well as the research hypotheses of Weisgarber et al.' original experiment, we now briefly summarize its methodology and results.

In terms of methodology, the original experiment consisted in four steps. Initially (1) the sample (n=257) has been divided in two groups by means of different instructions: in particular, the high-SD group (Group 1) had been told that the experiment was not anonymous and participants were required to insert their data, email and student ID; the low-SD group (Group 2), instead, had been assured that the answers collected were completely anonymous and required to provide honest and reliable responses. AS second step (2), both group were asked demographic questions and then the same health and wellness questions. The order of items was not randomized and the health and wellness questions were divided in four sections with four items per section (answers were provided as a five point Likert scale). After the

[5] E.g., response bias (Kumar, Kim, Valacich, Jenkins, & Dennis, 2021a) or fake identity detection (Monaro, Gamberini, & Sartori, 2017).

[6] In fact, many experimental studies proved that this domain is highly affected by the impact of SD bias, See e.g., (Levy et al., 2022).

target questionaire, a manipulation check was conducted through the question: "I felt my answers were anonymous.". Again, the answer was a five point scale ranging from 1. "Strongly disagree" to 5. "Strongly agree.". This manipulation check served to evaluate the effect of the different instructions (the manipulation) shown at the beginning of the experiment. Then (3), participants were required to complete Crowne & Marlowe' SD bias survey (MCSD), so as to be scaled in a consistent method in terms of their SD bias. At the end of the experiment (4), participants were informed about the actual purpose of the experiment and ensured that their responses were anonymous.

Having briefly exposed the methodology of the original study, we now presents its results and findings. Initially, the manipulation check proved that different instructions provided the expected effect: in fact, Group 1 exhibited a mean score of 3.45 while Group 2 scored 4.03. The statistical analysis conducted by the authors revealed that the difference was significant ($t\ (236) = -4.32$; $p < .001$; Cohen's $d = 0.56$): namely, Group 1 perceived less anonymity than Group 2. After checking that there was no relevant difference in the demographics and removing outliers,[7] the author used mixed-effect models with 'group' as fixed effect and 'participant' as subject adjustment to investigate the effect of anonymity vs non-anonymity on the four metrics measured (reaction time, mouse cursor deviations, mouse cursor speed, answer switches). The models indicated that H1 and H3 were supported by the data collected, whereas H2 and H4 were not. In Table 1, we report the table provided by the authors.

RH and Metric	**Result**	**p-value**
H1: Reaction time	Supported	.035
H2: Mouse cursor deviations	Not Supported	.324
H3: Mouse cursor speed	Supported	.001
H4: Answer switches	Not Supported	.423

Table 1: Results of the original study

[7] for more detailes on the method used for this latter procedure, See (Weisgarber et al., 2024), Sect. 5.2, p. 4678.

In addition to this analysis, the authors also performed a t-test to verify that the groups' mean scores in the MCSD questionnaire did not differ. Importantly, this was exactly the case.

These results showed that HCI dynamics can be used to detect social desirability bias in the context of health and wellness surveys. Furthermore, as the authors suggest, these findings are connected with three main practical implications: (1) HCI offer the chance to develop scalable methods to address SD bias in online surveys; (2) this framework gives insights on the participants decision-making process; (3) based on these results, it is possible to develop dynamic questionnaires that adjusts questions in accordance to the respondents SD bias level.

3 Motivations of the Study, Research Question and Hypotheses

In this section we briefly expose the motivation of our study and then clearly identify the research question and hypotheses that our experimental study is going to address.

3.1 Addressing limitations: SD bias in a new context

Although Weisgarber et al.' analysis represents a solid piece of research and contributes to social studies by suggesting a novel method to track SD bias, the authors themselves acknowledge that their experiment suffers from various limitations – principally four. Initially (1), their analysis is conducted on a specific demographic group: students of a north American university, primarily aged between 19 and 22. It may be that wider and most diverse population do not show the same behavior, and thus that the results of the study are not generalizable. A second core limitations (2) lies in the fact that the analysis is conducted on a specific topic (health and wellness questions). In order to generalize results and be able to argue for strong effects and correlations, the analysis should be replicated on other sensitive topics affected by SD bias responses. Furthermore (3), authors also acknowledge that there is a need for analyses to look for alternative explanations rather than the

effect of SD bias on the four metrics measures. Ultimately (4), the original study is constructed using a between subject design, but these kinds of behaviors can vary considerably across different subjects. Therefore, it may be useful to conduct with subject analyses to address this problem.

Our experimental analysis aims at addressing the second (2) core limitation listed by Weisgarber et al.: this means that we basically replicate the original study, but focus on a different topic than health and wellness. In particular, we run our analysis on AI chatbot interaction and privacy. It is no coincidence that Weisgarber et al. themselves list 'cybersecurity policy compliance' among the suggested fields to replicate their analysis on: researchers are well aware that self-reported data about cybersecurity and privacy policy compliance can likely be affected by SD bias[8]. We argue that the same is likely to happen as regards AI chatbot interaction domain. Therefore, our research question can be directly take from the original one, with one important modification, and be formulated as follows: *Do respondents' mouse movements differ in AI chatbot interaction and privacy surveys when social desirability is manipulated through two experimental conditions?* One could also think of our research question as follows: *Are the results found by Weisgarber et al. truly generalizable? Or, on the contrary, they are confined in a specific domain?*

3.2 Research hypotheses

Having clarified the motivation of our study and its main research question, we now present our research hypotheses. Briefly, we follow the footstep of the original study and hence hypothesize that respondents that are prompted to be less affected by SD bias (by means of anonymity and emphasis on honesty) will be less reflective and more impulsive in their decision-making process. On the contrary, respondents that are prompted to be more affected by SD bias (through non-anonymity and the absence of honesty requirements) will be more reflective and less impulsive while completing the survey.

This framework guides us in formulating the following three research hypotheses:

H1. *Respondents in the higher social desirability group will exhibit longer response times compared to the respondents in the lower social desirability group.*

H2. *Respondents in the higher social desirability group will exhibit slower mouse cursor speeds compared to the respondents in the lower social desirability group.*

H3. *Respondents in the higher social desirability group will exhibit more answer switches compared to the respondents in the lower social desirability group.*

Basically, our research hypotheses correspond to *H1, H3* and *H4* from the original study.

These hypotheses are used to investigate the activation of cognitive control (and hence the effect of SD bias) by means of HCI dynamics.

4 Methodology

The methodology of our experiment aims at replicating the original study as accurately as possible. This section summarizes such methodology, which is also exposed in Figure 1.

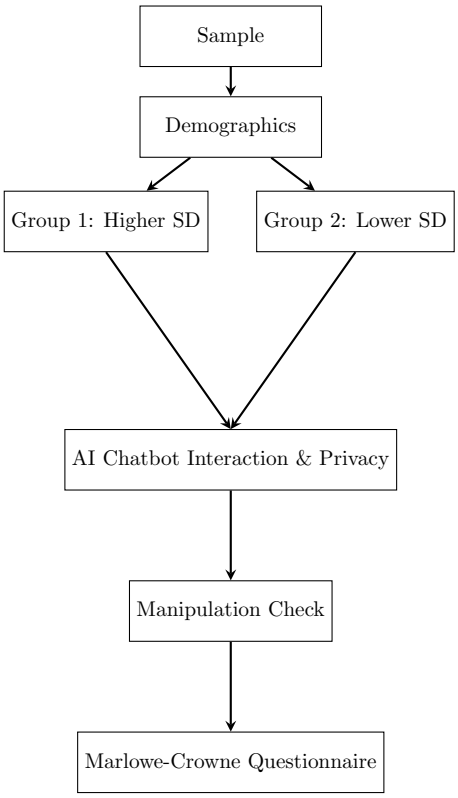

Figure 1: Methodology of the experiment

[8]E.g., Simonet & Teufel (Simonet & Teufel, 2019) acknowledge this in their experimental research.

4.1 Sample

The survey was sent to 52 subjects, but was completed by only 39 participants. 56.4% of the sample were males, and most of the participants were young people (only 25.6% were over 30). In terms of nationality, the most populated group is Italians (46.2%), but the combined number of participants from the rest of nationalities (including Dutch, Brazilian, and Spanish) accounts for the majority (53.8%). As regards job position, the largest group is students (53.8%), while he remaining 46.2% are workers (with most of them coming from the cybersecurity field). The sample of our analysis, hence, is more diverse than the original study; however, due to the small size of the population analyzed, we do not think that this sample is enough to address the first limitation considered by the authors (1).

4.2 Demographic and Instructions

The sample has been divided randomly in two groups (high-SD and low-SD). In particular, the high-SD group was asked, during the set of standard demographic questions, to insert full name and email at the beginning of the questionnaire, and received the following instructions before starting with the survey:
≪Your answers to this survey **will be recorded** and will be paired with the information previously provided on the "Personal information" section for later statistical analysis.≫.
The low-SD group, instead, was not asked to insert name nor email and was provided with the following instructions:
≪We wish to emphazise that you answers are going to be **anonymous**. Also, one of the most important aspects of surveys is to gather accurate information, so please answer the following questions **honestly**.≫.

4.3 AI chatbot interaction and privacy survey

After receiving the instructions exposed above, the two groups were introduced to the task by means of a short trial section consisting of only three items. Then, they were presented the target questionnaire. Our AI chatbot interaction and privacy questionnaire consists in two section (AI chatbot interaction – privacy and cybersecurity) with four items per section. An item of the first section, e.g., was: ≪How often do you ask chatbots for advice in topics related to your personal life (e.g. relationships, emotions, feelings)?≫. All the answers were provided with the same five point Likert scale ranging from 'Never' to 'Very often'. All in all, the survey consisted in eight experimental items, two control items and three fillers. The complete list of the questionnaire items is to be found in Appendix.[9] The items' order was the same for each participant, no matter the participant's group.

4.4 Manipulation Check

After completing the target survey, the same manipulation check adopted by Weisgarber et al. has been presented. The subjects were asked to rate their agreement to the statement 'I felt my answers were anonymous' on a scale from 1 = 'Strongly disagree' to 5 = 'Strongly agree'.

4.5 MCSD questionnaire

So as to have a measure of the participants SD bias, we employed the SD bias scale proposed by Crowne & Marlowe. However, in order to reduce the length of the experiment, we opted for the shorter 13 items version compared to the original one used in the original study (33 items).[10] At the end of the questionnaire, participants were ensured that their data were actually anonymous and acknowledged about the real purpose of the study.

4.6 Measures

As for metrics measured, of course we tracked participants mouse cursor movements and measured the three metrics required to address our research hypotheses. In particular, response time for H1, mouse cursor speed for H2 and finally answer switches for H3. In Table 2, we put the description of the three metrics.

In order to calculate mouse-tracking metrics such as mouse cursor speed, we employed the formatting method suggested by PCIBEX.[11]

[9] See Appendix A. Here the items are reported in the same order of the experiment.

[10] In particular, we employ the version C of the questionnaire, proposed by Crowne & Marlowe themselves as the most reliable short version of the scale((Crowne & Marlowe, 1960).

[11] See https://www.pcibex.net/wiki/mousetracker-

Metric	Description
Reaction time (ms)	The time spent by the subject in responding to target questions.
Mouse cursor speed (px/ms)	The average speed of the mouse cursor while responding to target questions.
Answer switches	The number of times that the subject changed answer during target questions.

Table 2: Description of the metrics measured

5 Results

Initially, we eliminated subjects that did not provide reasonable answers in the control items. As an example, one control item was: 'How often do you publicly share your online banking credentials on social media or public forums?'. It seems reasonable that if some subjects responded something as 'Often' or 'Really often', they are likely to either not understand the questions of the questionnaire or providing random answers without even reading the questions.

We must acknowledge an important difference between the original study and the present one in the statistical analysis. In the original study the authors calculated the mean (M) and median absolute deviation (MAD) for each metric, and deleted from the dataset data points greater than M + 3 * MAD.[12] We employed the same method as regards response time and average mouse speed, but not for answer switches. This is due to the fact that almost every subject did not show any answer switches. Hence, removing outliers would invalidate our results by deleting too many subjects.

5.1 Manipulation Check

To ensure that our manipulation instructions were effective, we performed a t-test on the mean scores of the manioulation check for both groups. The results show that our manipulation was effective, since the mean score difference between the two groups is statistically significant (p-value = 0.037).

element/.

[12]This approach was taken by from Kumar et al. (Kumar, Kim, Valacich, Jenkins, & Dennis, 2021b).

5.2 MCSD Questionnaire

As for the MCSD scale, we performed an independent t-test on the two groups' mean scores. Importantly, the test conducted does not suggest a significant difference between the two groups (p-value = 0.929). This means that participants from the two groups had, more or less, the same standard inclination to provide social desirable responses.

5.3 Metrics

For all the three metrics of our experimental research, we applied mixed-effect models with group as fixed effect and participant and question as random slope. This approach measures the effect of SD bias on the metrics and respects the assumptions of mixed-effect models. In fact, in each specific metric, we checked that the distribution of data was close to a normal one and, in case not, performed tailored transformations.

5.3.1 H1: Response Time

As for response time, we verified that the log transformation yielded the best with a normal distribution, as shown in Figure 2. Thus, we used logged RT as dependent variable in our mixed effect model. The model converged with the data, and the assumption of the model seem to be respected: in fact, its residuals are close to normal distribution (See Figure 3). However, the model did not predict any statistically significant effect of the variable group on logged RT (p-value = 0.464). This means that H1 is not supported.

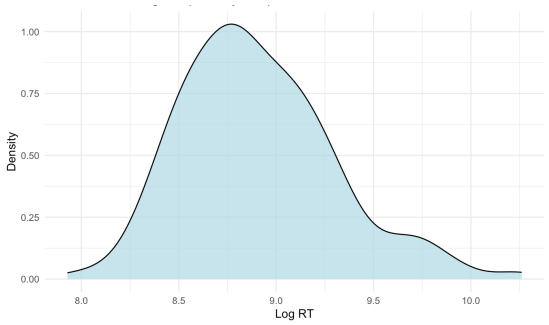

Figure 2: Plot of logged Response Time

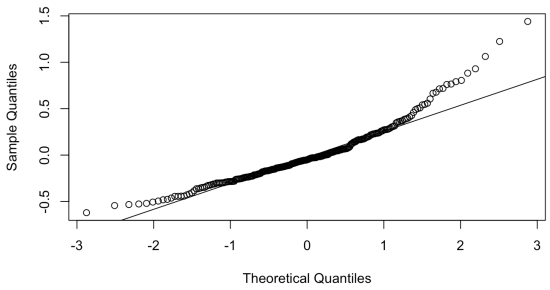

Figure 3: Residuals of the model for logged Response Time

5.3.2 H2: Mouse Average Speed

Also for mouse speed, log transformation is the most suitable option (see Figure 4). The model converged with the data and showed even more normally distributed residuals (Figure 5). No significant effect was found (p-value = 0.771) and, thus, also H2 is not supported.

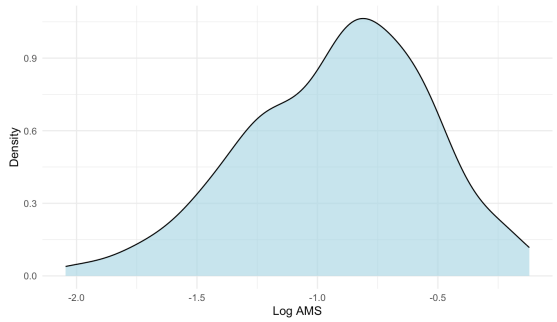

Figure 4: Plot of logged Average Mouse Speed

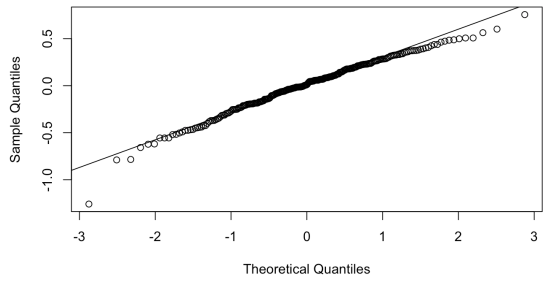

Figure 5: Residuals of the model for logged Average Mouse Speed

5.3.3 H3: Answer Swithces

As for the third metric, we employed a different type of model. In particular, we built a generalized mixed-effects model using poisson as regression family. This model is more suitable since the dependent variable is to be counted. As for the other metrics, the model converged but did not predict a significant effect on the dependent variable (p-value = 0.944): H3 is thus rejected.

All in all, none of the research hypothesis, as shown in Table 3, was supported.

RH and Metric	**Result**	**p-value**
H1: Reaction time	Not Supported	0.464
H2: Mouse cursor speed	Not Supported	0.771
H3: Answer switches	Not Supported	0.944

Table 3: Results of the experiment

6 Discussion

In this section, we expose the theoretical and practical implications of our results and also acknowledge its limitations.

6.1 Theoretical and Practical Implications

Our experimental study did not find any significant effect of SD bias on HCI dynamics tracked. Strictly speaking, this means that the answer to our research question is: *No*. Apparently the significant effects found in the original study were confined in a specific domain and are not generalizable. This could be due to the fact, as Weisgarber et al. point out, questions inherent to health and wellness are considered to be highly affected by SD bias. As regards privacy and policy compliance, some studies detect SD bias responses especially in working contexts; however, this phenomenon is not studied in the same extent as in the context of health questionnaires. Perhaps, HCI dynamics are useful to track SD bias (and in general the activation of cognitive control) in contexts were it has a considerable impact, but fail to produce valuable insights in fields were the effect is less sharp.

From a practical standpoint, this means that HCI dynamics are not yet been shown to produce insights on the participants' decision-making process and, thus, more studies are required to investigate this point. The hopes of Weisgarber et al. to produce a scalable method and be able to build dynamic questionnaires are

interesting and challenging; but they need to be based on more solid experimental results.

6.2 Limitations

Besides the results produced and our findings, we must acknowledge that our experimental study does suffer for various that could invalidate the strength of its results.

Initially (1) and most importantly, our sample was considerably smaller than the sample in the original study. The size of the sample had likely affected the findings of the experiment. Perhaps, running our same experiment on a larger sample could produce different results.

A second core limitation (2) lies in the fact that our experiment was conducted in English on non native speakers. Response time and Mouse speed are likely highly affected by the fact that respondents had to translate and understand each question before replying. In order to address this limitation, we suggest to replicate our experiment on native speakers.

The third limitation (3), connected to the second one, consists in the fact that we did not check for additional confunding variables that could have an effect on HCI dynamics.

Additionally (4), the need for a within subject design that could tackle the high variability of this data is yet to be addressed.

7 Conclusion

This study replicates the original experiment conducted by Weisgarber et al. on a different field (AI chatbot interaction and privacy), and thus aims at tracking SD bias through HCI dynamics such as mouse cursor movements. Our findings suggest that HCI dynamics fail to capture cognitive control activation and thus SD bias in the context of AI chatbot interaction and privacy surveys. This means that the method built in the original study is not generalizable and further analyses on other domains (such as political questions or social themes) are required to investigate the capability of HCI dynamics to detect SD bias in online survey.

However, we acknowledge that running our same experiment with a larger sample and on native speakers could likely modify the results obtained. Therefore, we suggest to replicate our study addressing its limitations in order to produce more valuable insights on this relevant topic.

References

Bhattacherjee, A. (2012). *Social science research: Principles, methods, and practices*. University of South Florida. Retrieved from https://repository.out.ac.tz/504/1/Social_Science_Research-_Principles_Methods_and_Practices.pdf

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. doi: 10.1037/h0047358

Hoffman, W. R., Aden, J. K., Barbera, D., & Tvaryanas, A. (2023). Self-reported health care avoidance behavior in u.s. military pilots related to fear for loss of flying status. *MILITARY MEDICINE*, *188*(3/4), e446. doi: 10.1093/milmed/usac311

Jenkins, J. L., Proudfoot, J. G., Valacich, J. S., Grimes, G. M., & Nunamaker, J. F. J. (2019). Sleight of hand: Identifying concealed information by monitoring mouse-cursor movements. *Journal of the Association for Information Systems*, *20*, 1–32. doi: 10.17705/1jais.00527

Kumar, M., Kim, D., Valacich, J., Jenkins, J., & Dennis, A. (2021b). Improving the quality of survey data: Using answering behavior as an alternative method for detecting biased respondents. In *Sighci 2021 proceedings* (p. 13). Retrieved from https://aisel.aisnet.org/sighci2021/13

Kumar, M., Kim, D., Valacich, J. S., Jenkins, J. L., & Dennis, A. R. (2021a). Improving the quality of survey data: Using answering behavior as an alternative method for detecting biased respondents. In *Sighci 2021 proceedings* (p. 13). Retrieved from https://aisel.aisnet.org/sighci2021/13

Levy, A. G., Thorpe, A., Scherer, L. D., Scherer, A. M., Drews, F. A., Butler, J. M., ... Fagerlin, A. (2022). Misrepresentation and nonadherence regarding covid-19 public health measures. *JAMA Network Open*, *5*(10), e2235837. Retrieved from https://doi.org/10.1001/jamanetworkopen.2022.35837 doi: 10.1001/jamanetworkopen.2022.35837

Monaro, M., Gamberini, L., & Sartori, G. (2017). The detection of faked identity using unexpected questions and mouse dynamics. *PLOS ONE*, *12*(5), e0177851. Retrieved from https://doi.org/10.1371/journal.pone.017785 doi: 10.1371/journal.pone.017785

Paulhus, D. L. (1988). *Balanced inventory of desirable responding* (Vol. 41). Acceptance and Commitment Therapy Measures Package. doi: 10.1037/t08059-000

Schuessler, K., Hittle, D., & Cardascia, J. (1978). Measuring responding desirably with attitude-opinion items. *Social Psychology*, *41*(3), 224. doi: 10.2307/3033559

Simonet, J., & Teufel, S. (2019). The influence of organizational, social and personal factors on cybersecurity awareness and behavior of home computer users. In *34th ifip international conference on ict systems security and privacy protection (sec)* (pp. 194–208). Lisbon, Portugal. Retrieved from https://inria.hal.science/hal-03744309/document doi: 10.1007/978-3-030-22312-0_14

Steenkamp, J.-B. E. M., De Jong, M. G., & Baumgartner, H. (2010). Socially desirable response tendencies in survey research. *Journal of Marketing Research*, *47*(2), 199–214. doi: 10.1509/jmkr.47.2.199

Tappin, B. M., van der Leer, L., & McKay, R. T. (2023). The heart trumps the head: Desirability bias in political belief revision. *Journal of Experimental Psychology: General*. doi: 10.1037/xge0001234

Weinmann, M., Valacich, J. S., Schneider, C., Jenkins, J. L., & Hibbeln, M. (2022). The path of the righteous: Using trace data to understand fraud decisions in real time. *MIS Quarterly*, *46*(4), 2317–2336. doi: 10.25300/MISQ/2022/17038

Weisgarber, P. A., Valacich, J. S., Jenkins, J. L., Kim, D., & Kumar, M. (2024). Detecting social desirability bias with human-computer interaction: A mouse-tracking study. In T. X. Bui (Ed.), *Proceedings of the 57th annual hawaii international conference on system sciences* (pp. 4673–4682). Honolulu: IEEE Computer

Society. Retrieved from https://aisel.aisnet.org/hicss-57/in/hci/4/

Appendix A: Items List

- Experimental item 1: "How often do you share personal and sensitive information with chatbots (e.g. sexual orientation, healthcare data)?".
- Experimental item 2:"How often do you ask chatbots for advice in topics related to your personal life (e.g. relationships, emotions, feelings)?".
- Filler 1: "How often do you see news or articles about artificial intelligence in your daily life (e.g., on social media, news websites, or TV)?".
- Experimental item 3: "How often do you chat with chatbots even if you don't need anything from them (e.g. just to chat)?".
- Experimental item 4: "How often do you rely on the answers of chatbots without checking them?".
- Filler 2: "How often do you receive suspicious emails or messages that might be phishing attempts?".
- Control item 1: "How often do you send and receive messages or emails?".
- Experimental item 5: "When you have to create a password for a new account, how often do you use strong passwords (different from any other password you used in the past, with many special characters and really long)?".
- Experimental item 6: "How often do you accept all the cookies (including the non-necessary ones) while browsing the internet?".
- Control item 2: "How often do you publicly share your online banking credentials on social media or public forums?".
- Experimental item 7: "How often do you read (even only partially) the privacy policy of a website before accepting it?".
- Filler 3: "How often do you watch tutorial videos or read guides to learn how to use new apps or digital tools?".
- Experimental item 8: "How often do you deploy 2-factor authentication (2FA)?".

Appendix B: Experiment code

We share all the codes of our experiment through a public github repository. In particular, the github depository includes:

- The code(s) used to produce the online survey, both for the high-SD group and the low-SD one.
- The csv file with all the data produced.
- The R.md file with the statistical analysis.

The git-hub repository is to be found in this link: https://github.com/DiegFari/Experiment-Codes.