RESOURCE and TIPS SHEET SUMMARY
WEB-BASED SURVEYS: BATTLING THE BOTS

On June 11, 2024, the Department of Physical Therapy Research Committee hosted a Rehabilitation Research Roundtable entitled: *Web-Based Surveys: Battling the Bots and Other Trials and Tribulations of Web-Based Survey Research in the Rehabilitation Sciences*.

The aim of this session was to profile and exchange knowledge within the Department of Physical Therapy on innovations and lessons learned in web-based survey research, specifically pertaining to issues with web-based survey recruitment (social media), data collection and responses by bots, AI-generated responses, and other threats to validity and to highlight key considerations for engaging in web-based survey research in the future.

The session included seven panelists, from the Department of Physical Therapy, Rehabilitation Sciences Institute and Department of Computer Sciences at the University of Toronto. Approximately 35 faculty, trainees, students and staff attended the event. This resource provides a summary of the 'tips for consideration' when conducting web-based surveys followed by resources that emerged from the session.

Tips for Consideration when conducting Web-Based Survey Research

Survey Questionnaire Development

- During survey design, build in strategies to identify suspicious or fraudulent activity within the questionnaire itself (e.g., CAPTCHA, seriousness checks, trick questions; email verification, validation and attention checks).

- Embed duplicate questions within the questionnaire or ask the same question in a slightly different way (e.g. including a question about age at the beginning and end of the survey instrument) – if responses differ this may indicate a potentially invalid response.

- If administering a survey questionnaire at multiple time points, ask for respondents to provide the same email address, so you can link responses, and check for consistency across survey questionnaires.

- Use multi-modal CAPTCHA. CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) are security measures put in place on places like websites to deter bot activities. It usually asks people to type in the weirdly shaped words in an image or select images that include prompted objects (e.g., traffic lights). There are tools that allow the bots to bypass the standard CAPTCHA, using AI. However, existing AI tools are comparatively underperforming for other modalities of data, such as videos. Therefore, in place for the standard CAPTCHA to deter the bot activities, a video that is easy to distinguish for humans (e.g., a video of an animal eating corn), with a multiple-choice question can help deter many of the bots now.

- Use more trustworthy recruitment platforms such as Prolific. Platforms like Amazon Mechanical Turk are known for higher noise rates (which can include bot activity), but there are platforms that have been shown to have lower noise rates. The recruitment procedure, therefore, could be improved by choosing the right recruitment strategy, such as the platform, to help reduce the bot activity in responses.

- Consider whether respondents will be compensated (e.g. e-gift card token of appreciation) and whether the compensation may incentivize potentially fraudulent responses in order to obtain compensation. If providing a token of appreciation, consider the amount in relation to the amount of time required to participate in the study.

- Enable metadata collection: Some platforms, like Qualtrics require you to manually turn on the metadata gathering, but there are many resources online that show how to record metadata and see it.

- Make sure to check if the platform you are using needs manual enabling of metadata collection.

Recruitment

- Be cautious about posting flyers on social media that state how much participants will be compensated (if token of appreciation is provided for participation) as this can attract invalid respondents.

- If collaborating with community organizations who are involved in recruitment, recommend that collaborator organizations recruit by circulating study materials internally for recruitment (i.e., monthly newsletters) and minimize posting on public social media platforms.

- Inspect your emails for signs of spam, such as a suspicious email address (first name + last name + string of numbers), possibly absent subject line, generic or absent greeting, poor grammar or spelling, and lack of clinic/hospital signature.

- Rather than openly posting the survey link, or including the link on recruitment materials, recruitment materials may ask interested individuals who meet study eligibility criteria to email the research team expressing their interest in the study. A member of the research team would then administer the survey link.

- Track your response rate early on. Response rates that exceed those published for studies using similar procedures can indicate fraudulent responses.

Data Cleaning

- Consult with the research team to obtain input on the proposed process for addressing potentially fraudulent data.

- The REAL Framework (Reflect, Expect, Analyze, Label) is a process in which to help guide identifying potential fraudulent responses to web-based surveys and make decisions about the inclusion or exclusion of responses based on level of suspicion.

- Create pre-data collection and post-data collection data quality checklists before final survey distribution. Some items to add on checklists:
  - Check time stamp (survey start and end time)
  - Time when survey completed (check for odd hours)
  - Valid email address
  - Check open ended responses asking for feedback if included in survey questionnaire for sensible responses.
  - Ensure no foreign characters in any text boxes

- Response time: Scan for multiple responses that have same time stamps (questionnaire initiation and completion end at the exact time down to the second or millisecond); and check for responses where the completion time is less than the amount of time estimated as needed to complete the questionnaire (information from the pilot process). This may indicate potential bot responses.

- Geographical Location & Time Zone: In some survey software, the browser automatically captures the time zone the responder is in. Research teams can filter out potential bots by ensuring the time zone that the responder is in matches their self reported location

- For post-processing text and filtering out AI generated responses (for open text responses): Input the survey question in the text box on ChatGPT and gather the output responses. All large language models are trained by maximizing ROUGE or BLEU scores (which are automatic metrics used to evaluate generated text quality). Calculate the ROUGE and BLEU score of the output from ChatGPT and responses from the online survey. Filter out the ones that have a score similar to the ROUGE and BLEU score of the ChatGPT output responses. Note, this will require some coding.

- GPTZero is an AI detector that is commonly used to detect AI generated content. Run GPTZero or similar AI detector scripts on the responses to detect AI generated text. Note, this will require some coding if you wish to automate the process.

- Document your data cleaning process carefully as you may need to repeat the process in future survey administrations and/or you will need to report on your procedures in publications and presentations.

Additional Considerations

- Be sure to keep the Research Ethics Board (REB) apprised of any issues that may arise in the form of annual renewals, or amendments. This is helpful so that they can continue to work on guidance for engaging in web-based survey research.

- Include in the study protocol (REB protocol) the steps to mitigate bot activity, and the steps in which the team will use in order to clean the data if a link is compromised. Having

a clear and consistent strategy which rationale for reasons why you included or excluded a response is imperative when writing up study findings.

- Consult with people who have experience with web-based survey research before survey development and ask them for feedback.

RESOURCES

Aday LA, Cornelius LJ. Designing and conducting health surveys: a comprehensive guide. John Wiley & Sons; 2011 Jan 20. https://www.wiley.com/en-us/Designing+and+Conducting+Health+Surveys%3A+A+Comprehensive+Guide%2C+3rd+Edition-p-9781118046678

Bonett S, Lin W, Sexton Topper P, Wolfe J, Golinkoff J, Deshpande A, Villarruel A, Bauermeister J. Assessing and Improving Data Integrity in Web-Based Surveys: Comparison of Fraud Detection Systems in a COVID-19 Study. JMIR Form Res. 2024; 8: e47091. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10818231/

Cook D, Heath F, Thompson RL. A meta-analysis of response rates in web or internet-based surveys. *Educational and Psychological Measurement*. 2000: 60 (6): 821-836. https://doi.org/10.1177/00131640021970934

Douglas BD, Ewell PJ, & Brauer M. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE*. 2023; 18(3), e0279720. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0279720

Equator Network. Enhancing the QUAlity and Transparency Of health Research. https://www.equator-network.org/reporting-guidelines/improving-the-quality-of-web-surveys-the-checklist-for-reporting-results-of-internet-e-surveys-cherries/

Equator Network. A Consensus -Based Checklist for Reporting of Survey Studies (CROSS) https://www.equator-network.org/reporting-guidelines/a-consensus-based-checklist-for-reporting-of-survey-studies-cross/

Eysenbach G. Improving the quality of web surveys: The checklist for reporting results of internet e-surveys (CHERRIES). *Journal of Medical Internet Research*. 2004; 6(3) :e34. https://doi.org/10.2196/jmir.6.3.e34 and Correction: https://www.jmir.org/2012/1/e8/

Eysenbach G, Wyatt J. Using the internet for surveys and health research. *Journal of Medical Internet Research*. 2002;4(2) :e13. DOI: 10.2196/jmir.4.2.e13. https://pubmed.ncbi.nlm.nih.gov/12554560/

Fan W, Yan Z. Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior*. 2010. 26:132-139. https://doi.org/10.1016/j.chb.2009.10.015

Goodrich B, Fenton M, Penn J, Bovay J, Mountain T. Battling bots : Experiences and strategies to mitigate fraudulent in onlnie surveys (2023). *Applied Economic Perspectives and Policy*. March 14, 2023. https://doi.org/10.1002/aepp.13353

Gupta N, Rigotti L, & Wilson A. The experimenters' dilemma: Inferential preferences over populations. Economics; General Economics. *arXiv preprint arXiv*:2021; 2107.05064. https://arxiv.org/abs/2107.05064. https://doi.org/10.48550/arXiv.2107.05064

Griffin M, Martino RJ, LoSchiavo C, Comer-Carruthers C, Krause KD, Stults CB, Halkitis PN. Ensuring survey research data integrity in the era of internet bots. *Quality and Quantity*. 2022. 56: 2841-2852 https://doi.org/10.1007/s11135-021-01252-1

Johnson MS, Adams VM, & Byrne J. Addressing fraudulent responses in online surveys: Insights from a web-based participatory mapping study. *People and Nature*. 2024; 6: 147–164. https://doi.org/10.1002/pan3.10557

Lawlor J, Thomas C, Guhin AT, Kenyon K, Lerner MD, UCAS Consortium, Drahota A. Suspicious and fraudulent online survey participation: Introducing the REAL framework. *Methodological Innovations*. 2021; 14(3); https://doi.org/10.1177/20597991211050467 https://journals.sagepub.com/doi/10.1177/20597991211050467

Levi R, Ridberg R, Akers M, Seligman H. Survey Fraud and the Integrity of Web-Based Survey Research. *American Journal of Health Promotion*. 2022; 36(1):18-20. doi:10.1177/08901171211037531. https://pubmed.ncbi.nlm.nih.gov/34372667/

O'Brien KK, Solomon P, Worthington C, Ibanez-Carrasco F, Baxter L, Nixon SA, Baltzer Turje R, Robinson G, Zack E. HIV, Health and Rehabilitation Survey Catalyst Team. Considerations for conducting web-based survey research with people living with human immunodeficiency virus using a community-based participatory approach. *Journal of Medical Internet Research*. 2014; Mar 13; 16(3); e81. https://pubmed.ncbi.nlm.nih.gov/24642066/

Pratt-Chapman M, Moses J, Arem H. Strategies for the Identification and Prevention of Survey Fraud: Data Analysis of a Web-Based Survey. *JMIR Cancer*. 2021. 7(3): e30730. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8325077/

Shaw TJ, Cascalheira CJ, Helminen EC, Brisbin CD, Jackson SK, Simone M, Sullivan TP, Batchelder AW, Scheer JR. 2024. *PsyArXiv Preprints*: https://osf.io/preprints/psyarxiv/gtp6z https://doi.org/10.31234/osf.io/gtp6z

Storozuk A, Ashley M, Delage V, Maloney EA. Got Bots? Practical recommendations to protect online survey data from bot attacks. *The Quantitative Methods for Psychology*. 2020. 16(5): DOI: 10.20982/tqmp.16.5. p472. Available at: https://www.tqmp.org/RegularArticles/vol16-5/p472/p472.pdf

Teitcher JEF, Bauermeister JA, Hoefer CJ, Miner MH, Klitzman RL. Detecting, Preventing, and Responding to "Fraudsters" in Internet Research: Ethics and Tradeoffs. *J Law Med Ethics*. 2015 Spring; 43(1): 116-133. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4669957/. https://doi.org/10.1111/jlme.12200

Wang J, Calderon G, Hager ER, Edwards LV, Berry AA, Liu Y, Dinh J, Summers AC, Connor KA, Collins ME, Prichett L, Marshall BR, Johnson SB. Identifying and preventing fraudulent

responses in online public health surveys: Lessons learned during the COVID-19 pandemic. *PLOS Global Public Health*. 2023; 3(8): e0001452. doi: 10.1371/journal.pgph.0001452 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10446196/

Xu Y, Pace P, Kim J, Iachini A, Bailey King L, Harrison T, DeHart D, Levkoff SE, Browne TA, Lewis AA, et al. Threats to online surveys: recognizing, detecting and preventing survey bots. *Social Work Research*. 2022.26(4); 343-350. https://doi.org/10.1093/swr/svac023

Zhang Z, Zhu S, Mink J, Xiong A, Song L, & Wang G. Beyond bot detection: combating fraudulent online survey takers. *ACM Web Conference*. 2022; 699-709. https://dl.acm.org/doi/pdf/10.1145/3485447.3512230