The Mode of Testing and Learning Outcomes: Evidence from In-Person and Phone Tests

Khandker Wahedur Rahman[1] Marjan Hossain[2]

May 25, 2023

Extended Abstract

Phone-based assessments gained prominence during the COVID-19 pandemic as an alternative to in-person assessments and as a consequence of growing concerns over monitoring children's learning during this extended period of school closures (Sobers, Anicet, Tanoh, Akpe, Ball and Jasińska, 2021). As a result of social distancing protocols during the pandemic, phone-based assessments were employed to measure the increase in learning levels on account of these interventions. This required the assessments to be adapted from their in-person versions, ensuring their administration across the telephone.

The implications of survey mode and its design characteristics on response distributions have been extensively documented (Biemer and Lyberg, 2003; De Leeuw et al., 2005; Klausch, Hox and Schouten, 2013; Schork, Riillo and Neumayr, 2021; Wagner, Schroeder, Piskorowski, Ursano, Stein, Heeringa and Colpe, 2017). Additional research investigating measurement equivalence across different survey modes (Ansolabehere and Schaffner, 2014; Zager Kocjan, Lavtar and Sočan, 2022; Zeglovits and Schwarzer, 2016) and the associated mode selection and measurement effects on responses (Shino, Martinez and Binder, 2022; Szolnoki and Hoffmann, 2013; Vandenplas, Loosveldt and Vannieuwenhuyze, 2016; Vannieuwenhuyze and Loosveldt, 2013), reveals that the survey mode can influence the quality of data collected, contingent on the type of outcomes examined. Findings indicate that social desirability biases in responses to alcohol consumption(Hox, de Leeuw and Klausch, 2017), health status (Hox et al., 2017), crime reports (Fong and Williams 2011), and political beliefs (Klausch et al., 2013; Zager Kocjan et al., 2022) are lower in self-administered modes (mail and web surveys) than in interviewer-administered modes (in-person and telephone surveys).

Test scores measured using phone-based surveys, similarly, can differ from if the tests were administered in-person to the same students; hence, the tests may not necessarily be comparable without knowing the standard difference between the mode of administration. Although face-to-face surveys incur a higher cost than mail, web, and telephone surveys, it is still the most preferred mode among researchers and is regarded as the benchmark for good data quality (Biemer and Lyberg, 2003). Despite the fact that phone assessments are more cost-effective and desirable in resource-constrained contexts or crisis scenarios (Angrist, Bergman, Evans, Hares, Jukes and Letsomo, 2020; Crawfurd, Evans, Hares and Sandefur, 2021; Rodriguez-Segura and Schueler, 2022), any potential large-scale shift from in-person to phone-based assessments necessitates additional research to produce equivalent measures.

In this paper, we establish the equivalence between a phone-based survey and an in-person

[1]University of Oxford and BRAC Institute of Governance and Development, BRAC University.
[2]BRAC Institute of Governance and Development, BRAC University.

survey. We administer a standardized test on 1,396 Bangladeshi children aged 6-18 where each child took the test twice: in-person and over the phone. To attenuate the potential effect of the order in which survey modes are administered on learning levels, we randomly assigned the sampled households into two groups. Households in group A were given phone surveys before being visited in person, while households in group B were given phone surveys after being visited in person. Children in all of the households received the test twice with a gap of 7-14 days kept between each test mode. We then compare the deviation of the phone survey scores from the in-person survey (the benchmark) to generate an equivalence measure of the two.

The test module was designed to measure foundational levels of literacy/numeracy consisting of all items from the Bangladeshi adaptation of the Annual Status of Education Report (ASER) test, along with some additional questions used in Crawfurd et al. (2021) which were modified for our study context. The ASER instrument for Bangladesh contains three modules: Bangla, English, and Mathematics. Our testing approach was tailored to capture a child's maximum proficiency level. The child was presented with a series of increasingly difficult questions and was allowed to continue until they could no longer provide correct responses.

To ensure consistency in the testing methods for both in-person and phone evaluations, we adapted our methods to accommodate each mode of the survey. Visual cues were provided for the ASER items in the test instrument, and bonus questions were administered verbally in both cases. For in-person tests, children were given handouts containing the ASER questions to read and solve, while questions were texted to students for phone assessments. While enumerators were able to discern when the child was ready to answer during in-person testing, this was relatively more challenging to do so over the phone. To address this, we implemented specific protocols for phone evaluations. Enumerators first asked the respondent if they could open and read text messages and then sent questions by text message. The test began only after the child confirmed that they had received the text messages and were prepared to start answering. These measures were used exclusively for administering the ASER test items and were not necessary for the bonus questions, which were narrated to the child over the phone. In both survey modalities, testing was subject to a time constraint. Each question had a 4-minute time limit (except for the Bangla short story, which had a limit of 5 minutes). Before starting the test, the child was informed of this time limit and also reminded before each question.

We find that, on average, students score more on the phone survey. We see that students score between 0.07-0.17 SD more on the phone survey than on the in-person survey. We further find that this trend remains similar for both boys and girls; however, the overall difference for boys is smaller than that for girls. Students spent less time completing the phone-based tests as well. We find that, on average, the students spent about ten fewer minutes completing the phone-based test.

This paper documents evidence on how comparable are phone-based tests with in-person tests. With the breakout of the pandemic, the importance of phone-based tests grew, but there is little known about how comparable they are with in-person tests. This issue of

comparability is of utmost importance that will likely influence inference that are based on phone-based tests, and, hence, has policy implications. In addition, we adopted several innovations to administer phone tests in a low-technology setting that are important contribution to the literature.

References

Angrist, Noam, Peter Bergman, David K. Evans, Susannah Hares, Matthew C. H. Jukes, and Thato Letsomo (2020) "Practical lessons for phone-based assessments of learning," *BMJ Glob Health*, 5 (7), e003030, 10.1136/bmjgh-2020-003030.

Ansolabehere, Stephen and Brian F Schaffner (2014) "Does survey mode still matter? Findings from a 2010 multi-mode comparison," *Political Analysis*, 22 (3), 285–303.

Biemer, Paul P. and Lars E. Lyberg (2003) *Introduction to Survey Quality*, Wiley Series in Survey Methodology, Hoboken, NJ, USA: John Wiley & Sons, Inc. 10.1002/0471458740.

Crawfurd, Lee, David K. Evans, Susannah Hares, and Justin Sandefur (2021) "Teaching and Testing by Phone in a Pandemic,"Technical report, Center for Global Development, Washington, DC.

De Leeuw, Edith D et al. (2005) "To mix or not to mix data collection modes in surveys.," *Journal of official statistics*, 21 (5), 233–255.

Hox, Joop, Edith de Leeuw, and Thomas Klausch (2017) "Mixed-Mode Research," in *Total Survey Error in Practice*, 511–530: John Wiley & Sons, Ltd, 10.1002/9781119041702.ch23, Section: 23 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119041702.ch23.

Klausch, Thomas, Joop J. Hox, and Barry Schouten (2013) "Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions," *Sociological Methods & Research*, 42 (3), 227–263, 10.1177/0049124113500480.

Rodriguez-Segura, Daniel and Beth E. Schueler (2022) "Can learning be measured by phone? Evidence from Kenya," *Economics of Education Review*, 90, 102309, 10.1016/j.econedurev.2022.102309.

Schork, Joachim, Cesare A. F. Riillo, and Johann Neumayr (2021) "Survey Mode Effects on Objective and Subjective Questions: Evidence from the Labour Force Survey," *Journal of Official Statistics*, 37 (1), 213–237, 10.2478/jos-2021-0009.

Shino, Enrijeta, Michael D Martinez, and Michael Binder (2022) "Determined by Mode? Representation and Measurement Effects in a Dual-Mode Statewide Survey," *Journal of Survey Statistics and Methodology*, 10 (1), 183–202, 10.1093/jssam/smab012.

Sobers, Shauna-Marie, Konan Nana N'Goh Anicet, Fabrice Tanoh, Yapo Hermann Akpe, Mary-Claire Ball, and Kaja K. Jasińska (2021) "Is a Phone-Based Language and Literacy Assessment a Reliable and Valid Measure of Children's Reading Skills in Low-Resource Settings?," October, 10.35542/osf.io/ytvn4.

Szolnoki, Gergely and Dieter Hoffmann (2013) "Online, face-to-face and telephone surveys—Comparing different sampling methods in wine consumer research," *Wine Economics and Policy*, 2 (2), 57–66, 10.1016/j.wep.2013.10.001.

Vandenplas, Caroline, Geert Loosveldt, and Jorre T. A. Vannieuwenhuyze (2016) "Assessing

the Use of Mode Preference as a Covariate for the Estimation of Measurement Effects between Modes. A Sequential Mixed Mode Experiment," *methods, data, analyses*, 10 (2), 24, 10.12758/mda.2016.011, Number: 2.

Vannieuwenhuyze, Jorre T. A. and Geert Loosveldt (2013) "Evaluating Relative Mode Effects in Mixed-Mode Surveys:: Three Methods to Disentangle Selection and Measurement Effects," *Sociological Methods & Research*, 42 (1), 82–104, 10.1177/0049124112464868, Publisher: SAGE Publications Inc.

Wagner, James, Heather M. Schroeder, Andrew Piskorowski, Robert J. Ursano, Murray B. Stein, Steven G. Heeringa, and Lisa J. Colpe (2017) "Timing the Mode Switch in a Sequential Mixed-Mode Survey: An Experimental Evaluation of the Impact on Final Response Rates, Key Estimates, and Costs," *Social Science Computer Review*, 35 (2), 262–276, 10.1177/0894439316654611, Publisher: SAGE Publications Inc.

Zager Kocjan, Gaja, Darja Lavtar, and Gregor Sočan (2022) "The effects of survey mode on self-reported psychological functioning: Measurement invariance and latent mean comparison across face-to-face and web modes," *Behav Res Methods*, 10.3758/s13428-022-01867-8.

Zeglovits, Eva and Steve Schwarzer (2016) "Presentation matters: how mode effects in item non-response depend on the presentation of response options," *International Journal of Social Research Methodology*, 19 (2), 191–203, 10.1080/13645579.2014.978560, Publisher: Routledge _eprint: https://doi.org/10.1080/13645579.2014.978560.