Social media as a recruitment and data collection tool: Experimental evidence on the relative effectiveness of web surveys and chatbots*

Emily A. Beam

September 18, 2022

Abstract

Online technologies enable lower-cost, rapid data collection, but concerns about access and data quality impede their use in global research. I conduct a randomized experiment in the Philippines to test the effectiveness of web-form and chatbot surveys of K–12 teachers recruited through social media and compare their effectiveness with phone surveys of teachers recruited from a pre-existing frame. Chatbot surveys yield higher response rates and higher-quality data than web-form surveys in terms of missed question and item differentiation. The results suggest that chatbot responses match CATI responses on multiple dimensions of quality. Relative to CATI, online methods also yield higher rates or information disclosure on potentially sensitive topics, revealing substantially higher levels of distress among teachers. I show that social-media-based recruitment can be an attractive alternative for targeted sampling and that online surveys can be implemented effectively at a fraction of the cost of phone surveys.

*This study was supported by the UBS Optimus Foundation and benefited from the support of the Philippine Department Education Policy Research Division. Karisha Anne Cruz, Jed Dimaisip-Nabuab, and Rene Marlon Panti at Innovations for Poverty Action and provided outstanding project management and research assistance, and Nassreena Sampaco-Baddiri provided exceptional project leadership and support. I also thank the team of enumerators and supervisors at IPA-Philippines who made this project possible. Many thanks to Anne Fitzpatrick and Molly Offer-Westort for extremely useful feedback on the project and paper. The IPA Human Subjects Committee provided oversight for this project, "Philippines DepEd Needs Assessment," protocol #15695.

1 Introduction

Remote surveys in low- and middle-income countries can reduce data collection costs while expanding and transforming the questions researchers can answer when they can reach respondents rapidly across a wide geographic area. The rapid global expansion of mobile phone access, particularly over the last ten years, has enabled the rise of remote phone-based surveys in developing countries, primarily through computer-assisted telephone interviews (CATI) (Henderson and Rosenbaum, 2020). Remote methods, essential during emergencies when in-person data collection is difficult or impossible (Etang and Himelein, 2020), permit higher-frequency or lower-cost data collection in more general settings (Fafchamps and Minten, 2012; Dillon, 2012; Arthi et al., 2018; Garlick, Orkin and Quinn, 2020) and may be more reliable and safer when asking about sensitive topics (Ellsberg et al., 2001; Heise and Hossain, 2017; Assefa et al., 2022).

Among the portfolio of remote methods, online surveys permit new sampling options while allowing more flexibility in question design and implementation relative to phone or face-to-face surveys. Because of their reduced cost and rapid deployment, the use of web-form surveys has grown substantially in wealthy countries. The recent and ongoing rise in online access—currently, 63 percent of the world's population are internet users, a 17-percent increase since 2019 (ITU, 2021)—makes online surveys increasingly feasible for a growing number of global contexts, particularly when working with targeted populations with high rates of internet access (Lau et al., 2018; Rosenzweig et al., 2021).

Online surveying also permits the use of social media for remote recruitment and data collection. Recruiting survey respondents through social media can be cost-effective, particularly when the target population has a high rate of internet penetration (Rosenzweig et al., 2020; Pham, Rampazzo and Rosenzweig, 2019), and it can be particularly useful to recruit otherwise difficult-to-reach populations.[1] Chatbots deployed through social media

[1]For example, Jäger (2017) uses Facebook to recruit political activists in Thailand, as does Kapp, Peters and Oliver (2013) to recruit female smokers in the United States.

can survey respondents and deliver interventions (Rick, Goldberg and Weibel, 2019; Fulmer et al., 2018), providing a more interactive—and potentially more personal— platform than a traditional web-form survey (Kim, Lee and Gweon, 2019). Despite these potential advantages, the use of chatbot surveys in low- and middle-income countries has been rare,[2] and little is known about chatbot data quality relative to other remote methods, such as web-form surveys and phone surveys.

I implement a randomized experiment to measure differences in response rates and data quality of surveys conducted by web form and chatbot in a developing country context. Specifically, I survey 2,063 K–12 public school teachers in the Philippines about their personal well-being and experience with remote teaching during the COVID-19 pandemic. Smartphone ownership is near-universal among public school teachers (94 percent), making this context well-suited to compare the effectiveness of these methods.[3] I recruit the internet samples through targeted Facebook advertising, and participants are randomized into the two different survey modalities. I compare the response quality and content between the two online modalities, and I conduct a descriptive comparison with 1,229 CATI surveys with teachers, which is based on a sampling frame drawn from the full universe of public school teachers.

I report three main findings. First, social media recruitment yields full regional coverage of teachers, and teachers participating through online surveys are no more likely to have access to a smartphone or laptop than the overall population. Additionally, the underlying age distribution of online respondents is similar to that of the overall population of teachers. In contrast, phone survey participants are likely to be younger and more likely to own a cellphone or laptop than both the online sample and national population of teachers. Phone survey respondents are also less represented in very rural regions, while the online respondents, particularly those reached by chatbot, are overrepresented in metro-Manila.

[2]One exception is ongoing work by Offer-Westort, Rosenzweig and Athey (2021).

[3]Internet access, often via smartphone, is high across the Philippines. As of 2019, 70% of adults used the internet at least occasionally or owned a smartphone, with rates of 74% for those 30–49 and 94% for those 18–29 (Schumacher and Kent, 2020).

Second, I find that chatbot responses yield higher response rates and are of higher quality than web-form surveys on multiple dimensions. The chatbot completion rate, conditional on clicking on the invitation, was 48 percent, compared with 29 percent among web-form respondents, despite the survey taking longer to complete. This difference is driven by those who click on the web-form invitation but never begin the survey.

A key concern of web-form surveys is that respondents may reduce their cognitive load through "satisficing" (Krosnick, 1991), in which they provide a lower-effort satisfactory answer rather than expending additional effort to give the optimal answer. Indeed, chatbot surveys have a 12 percentage-point lower rate of straightlining in which respondents enter the same item choice (like "strongly agree") to a set of similarly structured items, relative to web-form surveys. The chatbot straightlining rate is statistically indistinguishable from the 24 percent rate among CATI. Chatbot surveys also reduce the share of questions skipped or answered with "don't know" relative to web-form surveys. Because the phone modality is not randomized, these differences could reflect differential selection into survey participation across modalities as well as the impact of the modality itself, although I find that the extent of selection on unobservables would need to be 1.5–2.4 times higher than selection on observables to fully explain these quality differences (Altonji, Elder and Taber, 2005; Oster, 2019).

Chatbot respondents are also more willing than web-form respondents to disclose potentially sensitive information: chatbot respondents report 0.4 standard deviations higher PHQ-4 scores (measuring depression and anxiety). Relative to CATI, online respondents appear more likely to disclose potentially sensitive information, measured along dimensions of mental health and general well-being during COVID-19, although I cannot causally isolate whether this is driven by modality or selection. Thus, the choice of sampling and survey method leads to meaningful differences in the overall measurement of teachers' experiences, in particular the measurement of mental health distress. Teachers surveyed online report 0.21 standard deviations higher PHQ-4 scores (for depression and anxiety) relative to teach-

ers surveyed by phone. And while 18 percent of teachers surveyed by phone say they are less able to balance work and life during the pandemic, 44–45 percent of online respondents say they are less able. Weighting the sample to reflect the national distribution of teachers does not affect this difference. Overall, these results are consistent with literature in high-income countries that finds web-form survey data is of lower quality (Fricker et al., 2005; Heerwegh and Loosveldt, 2008) but its less-personal nature reduces social desirability bias (Kreuter, Presser and Tourangeau, 2008), increasing respondents' willingness to share information about potentially sensitive topics.

Third, I document that the marginal costs of implementing an online survey are a fraction of phone survey implementation costs (less than $1 per respondent vs. $6.30 per respondent), and the higher completion rate from chatbot surveys makes them cheaper than web-form surveys on average. The low cost of online surveys, along with their ability to be implemented rapidly, provides new opportunities to collect information on a broader range of outcomes, generate higher-frequency data, and enable researchers with limited budgets to conduct larger-scale data collection. In cases where some face-to-face or CATI surveys are feasible, online surveys may be a useful way to reduce survey costs or increase response rates through mixed-mode data collection (De Leeuw, 2005) or to quickly pilot and conduct exploratory analyses.

This study demonstrates that social media recruitment can be an effective tool for sample generation in low- and middle-income country settings, particularly when researchers aim to target a particular sub-population. When surveying remotely, pre-existing sampling frames can yield high contact rates, but they are not always representative, if available at all (Henderson and Rosenbaum, 2020). Random-digit dialing may yield more representative samples (of phone users), but screening can bring substantial costs, and further targeting can be expensive. This study joins recent work by Pham, Rampazzo and Rosenzweig (2019) and Rosenzweig et al. (2020), which examine the extent to which Facebook advertising can

generate nationally representative samples in Mexico and Kenya.[4]

Additionally, this study complements previous work on the effectiveness of online surveys by measuring selection into participation and data quality in lower-income contexts, where barriers to usage may be higher. Consistent with studies in high-income countries, this paper suggests that web-form data may be of lower quality than CATI data. On the other hand, online surveys of either type appear to reduce social desirability bias (Kreuter, Presser and Tourangeau, 2008; Lee et al., 2019; Amaral et al., 2022), and chatbots excel particularly, which can be critical when investigating sensitive topics.

This paper also highlights that using chatbots remedies many observed weaknesses of web surveys in terms of data quality. This is in line with Kim, Lee and Gweon (2019), who find that chatbot surveys yield higher quality data in terms of non-differentiation, although this study has a larger sample, focuses on a wider range of data quality measures in a lower-income context, discusses the relationship between modality and the distribution of responses, and allows a comparison of online with CATI methods.[5]

The results of this study also provide insight into the potential role of remote surveys as a complement to or substitute for face-to-face surveys. The low cost of implementation, coupled with the relative effectiveness of online methods, means that online surveys have the potential to expand the nature and frequency of data collection. Lower-cost research methods have substantial equity implications by reducing barriers for researchers from low- and middle-income countries and early-career scholars.

[4]Both papers find substantial differences in the demographic characteristics of recruited samples, and Rosenzweig et al. (2020) finds that weighting can modestly reduce, but not eliminate, these differences.

[5]Specifically, Kim, Lee and Gweon (2019) conduct a randomized experiment with 117 adolescents in South Korea to measure the impact of modality (chatbot versus web-form) and conversational tone (casual vs. formal) on item differentiation, ease of use, and user enjoyment, also finding that chatbots create a more positive user experience.

2 Remote survey methods in low- and middle income settings

Phone-based and online survey methods can have multiple applications in low- and middle-income country settings, although their usefulness depends on the local context, target population, and researcher goals. Remote survey methods require that researchers can reach participants by phone or internet, which requires device and service access, reliable networks, and, in the case of internet or text-message-based surveys, literacy and technological familiarity. As cellphone and internet access has grown consistently worldwide, even among very low-income populations, the potential range of participants has also grown.

First, in target populations with even moderate rates of internet use, CATI and online surveys are particularly useful for low-cost testing and rapid piloting, for which researchers prioritize understanding potential patterns of responses over capturing a representative sample. Survey platforms such as ODK use standardized coding across CAPI, CATI, and web-form surveys, which allow online or phone-based piloting without additional programming costs. While most chatbot platforms require separate programming, I find that they yield higher response rates and higher quality data.

Second, both online and phone surveys can facilitate mixed-mode data collection (De Leeuw, 2005) to maximize response rates or reduce costs. They also permit higher-frequency data collection between in-person survey rounds, which is particularly important when investigating outcomes that are time-sensitive or subject to recall bias, like food consumption or daily business revenue.

Finally, in the case of sensitive questions, remote surveying may be preferable to in-person in LMIC surveying by reducing social desirability bias and increasing respondent privacy. Respondent privacy is especially difficult to obtain when households live in close quarters (Assefa et al., 2022), and it is particularly important when overheard answers could have negative repercussions for the respondent, such as with questions about women's empower-

ment and intimate partner violence (IPV) (Ellsberg et al., 2001; Heise and Hossain, 2017). Assefa et al. (2022) find that CATI yield higher disclosure rates of sensitive responses on topics around gender norms and IPV in Ethiopia. Additionally, the relative anonymity of answering by phone, and especially online, increases respondents' willingness to respond honestly on topics tied to their own self-presentation and desire for social approval (Krumpal, 2013). Among remote options, respondent literacy and comfort with technology will drive whether CATI or online surveying is more appropriate. For example, Park et al. (2021) find that respondents answering by audio-computer assisted self-interviewing (ACASI) reported higher rates of IPV, but their lack of familiarity with the technology meant that one-third of ACASI respondents failed to correctly answer objective screening questions.

Web-form and chatbot surveys have advantages relative to CATI and other phone-based methods. A common problem with CATI surveys in low- and middle-income contexts is dropped connections due to weak cellphone signals or poor connection quality. This can be exacerbated by bad weather, as well as if a respondent is interrupted and needs to continue the survey later. Scheduling suitable times by phone with respondents can be challenging, particularly for respondents who are only available outside traditional working hours. Online surveys enable respondents to select a convenient time to participate, and surveys can be resumed easily after losing connectivity. Chatbot surveys are particularly well-suited to interruptions, as they remain a chat message in the respondent's app, and the software enables implementers to send a "nudge" to encourage completion. Additionally, chatbots can be embedded in popular social media platforms like Facebook and WhatsApp, which may also be free to respondents through popular promotional packages or through zero-rating, in which a provider provides free access to a particular platform, waiving data usage costs (Rosenzweig et al., 2021). While this study recruits respondents through social media, integration with a platform like WhatsApp enables researchers to combine chatbots with pre-existing sampling frames.

3 Study Design

I conduct this study in the context of a survey of 3,292 K–12 teachers in the Philippines, conducted in partnership with Innovations for Poverty Action and the Philippine Department of Education. This survey aimed to identify the needs of teachers who were adapting to fully remote learning for the 2020–2021 school year. The first round of 1,331 online-only surveys took place in August 2020, just before the start of the delayed (remote) academic year. In December 2020, I conducted 1,229 surveys by phone and 732 surveys online (1,961 total) with an additional cross-section of teachers to understand how their needs had changed now that they had been teaching for four months. Figure 1 details the timeline of the full study.

To be eligible, participants had to be K–12 teachers. In the online version of the survey, this was stated in the recruitment materials. After the survey introduction but prior to consent, respondents were asked to confirm if they were K–12 teachers. Participation was not incentivized.

3.1 Online sample

Recruitment and randomization: I recruited online participants using a Facebook ad campaign (see Appendix Figure A.1). Facebook is an ideal recruitment platform given its high rate of penetration nationally (Rosenzweig et al., 2021). Additionally, nearly all teachers (88 percent) reported using Facebook Messenger to communicate with their students. The target audience fell into one of two categories: (1) people who listed the Philippine Department of Education as their employer; or (2) people who listed the Philippines as their country and either (a) listed teacher, elementary teacher, or high school teacher as their occupation, or (b) listed teacher as interests. Among those reached online, individuals were randomly selected to be surveyed by chatbot or web form (Figure 2).[6] The advertisements

[6]In the second round of surveying, I further randomized web respondents into two different platforms, SurveyCTO and Qualtrics, in order to determine whether aesthetic considerations affected completion rates. Because completion rates are equivalent (31 percent vs 29 percent in round 2), I separate platforms only when considering discussing completion rates, for which Qualtrics has more detailed information, and when

that participants received were identical except for the embedded survey link corresponding to the online survey modality.

Implementation: Appendix Figure A.2 shows the user interfaces for the chatbot and the web-form survey (SurveyCTO). The chatbot interface is fully integrated into Messenger, so the respondent interacts in the same way as a standard messaging exchange. The primary difference is that in the case of multiple-choice and yes-no questions, respondents also have the option to press or click on a button rather than typing their responses.

Specifically, the chatbot introduces the survey and asks whether the respondent would like to participate, which she can do by clicking or pushing on the appropriate button or by typing the associated number. Multiple-choice questions allowed respondents to press similar visual buttons. For open-response questions, respondents entered a number or word(s) using the keyboard or keypad on their device. In the case of invalid responses, such as typing a word instead of selecting a specific option, the chatbot responded that it could not understand the response and asked the respondent to try again.

The web-form survey link takes the respondent to a webpage in which the survey is introduced, and respondents click on response choices or enter a text or numeric response, as appropriate. The method of response, such as clicking on an answer choice or entering text, is held constant between the chatbot and web-form survey.

3.2 Phone sample

I first randomly selected three (3) schools per region and school level (elementary, junior high school, senior high school) from the universe of all public schools in the Philippines. The DepEd Central Office coordinated with the respective regional office or school division office to collect teacher contact information from this set of randomly selected schools. The Department of Education required that teachers agree to provide their contact information,

measuring costs.

introducing selection into the sample, although those who did not agree to share their information presumably would have been unlikely to participate in the survey. Upon obtaining teacher contact information for the set of selected schools, I randomly selected teacher-respondents, stratifying by region and school level. Surveys were conducted by trained enumerators working for Innovations for Poverty Action. They used SurveyCTO to encode responses during the phone survey.

4 Empirical Strategy

A prime concern with online surveys is that in addition to lower completion and response rates, respondents may put forth less effort relative to CATI or face-to-face surveys, leading to lower data quality and reduced statistical power (Heerwegh and Loosveldt, 2008; Heerwegh, 2009; Oppenheimer, Meyvis and Davidenko, 2009). Specifically, the elimination of interviewers and the structure of web-form surveys may encourage satisficing as a way to reduce effort (Fricker et al., 2005; Kim et al., 2019). However, online surveys could instead reduce satisficing by allowing respondents to select the time that works best for them, and the ability to reread questions could reduce cognitive burden and improve response quality (Fricker et al., 2005).

Satisficing can be detected through lower differentiation between items, such as when respondents "straightline" by providing the same answer to a set of similarly wording questions (i.e., always selecting "agree" on a series of Likert-scale questions). Additionally, satisficing can lead more respondents to answer "don't know" or skip questions.

Online surveys also may improve respondents' willingness to answer potentially sensitive questions because they are not speaking to an interviewer in person or over the phone. Previous studies find web-form surveys increase the likelihood of disclosure due to reduced social desirability bias (Kreuter, Presser and Tourangeau, 2008; Lee et al., 2019).

I measure the impacts of survey modality on response quality along four dimensions:

(1) item differentiation, or the extent to which respondents either avoid straightlining or use more elements of a point-scale when responding to a battery of similarly structured questions; (2) the likelihood that respondents answer that they "don't know" or skip a question; (3) the likelihood of extremely short surveys (less than 1.5 s.d below the mean duration, or less than 7 minutes); and (4) willingness to respond to potentially sensitive (mental health and well-being) questions.

To do so, I estimate the following model for each quality dimension:

$$y_i = \beta_0 + \beta_1 Online_i + \beta_2 Chatbot_i + \beta_3 R2_i + X_i'\beta + \epsilon_i$$

where y_i is the outcome of interest for respondent i, $Online$ is a binary variable equal to one for individuals surveyed by web survey or chatbot, $Chatbot$ is a binary variable equal to one for individuals surveyed by chatbot, such that phone survey respondents form the omitted category. $R2$ is an indicator for the second survey round.[7] I include a vector of individual-level covariates, X_i, which include day-of-week fixed effects along with demographic controls for teacher gender, age, position, school level, education level, and whether they have children under 18 living at home.

To see how these results correspond to the national population of teachers, I adjust the weights of the online sample and the phone sample to reflect external administrative records based on the distribution of teachers by gender and grade taught, and the distribution of teachers by region and laptop ownership. Because disaggregated data across all four dimensions were not available, I implement iterative proportional fitting to create weights that reflect the distribution of teachers in the two administrative data sets.[8] To compare sample characteristics across the two recruitment and collection methods, I generate weights

[7] During round 2, timing for the phone and the online survey differs by 2–3 weeks. In particular, the second online wave took place in late December, overlapping with holiday celebrations for many respondents, which could affect respondent attitudes and attentiveness. Adding a control for being surveyed during the winter break does not affect the magnitude nor significance of the results, somewhat reducing the concern that timing is driving the results.

[8] I use the `survwgt` package (Winter, 2018) in Stata. Appendix Tables A.3, A.4, A.5, and A.6 show that results are robust to alternative weights.

separately for the online data set and the phone data set.

5 Results

5.1 Response and completion rates

I first test the impact of survey modality on response and completion rates, as these factors have direct implications for the ability of the survey to reach the targeted population and for costs. I measure completion rates as the likelihood that a respondent completed a survey conditional on answering the phone or clicking on a Facebook advertisement to begin a survey. Chatbots yield substantially higher completion rates relative to web-form surveys, with rates of 48 percent versus 29 percent, respectively, as Table 1 shows. Phone survey completion rates, conditional on answering the phone, are much higher, at 90 percent. While the difference in completion rates between chatbots and web surveys reflect random assignment to treatment and identical measurement, the difference in completion rates between phone and online surveys also reflects differences in sample recruitment. The overall response rate for those in the phone sampling frame was 75 percent, which is higher than in many other CATI surveys (Henderson and Rosenbaum, 2020), likely reflecting that teachers had already agreed to provide their contact information and that as a relatively higher-earning population, they likely have fewer connectivity issues and may be less likely to change mobile numbers over time.

Because targeting is imperfect, online survey respondents are screened only after clicking on the advertisement, and not all consent. Completion rates, once started, are only modestly higher for chatbot versus web-form surveys. Among those who consent to participate in online surveys, 76 percent of chatbot respondents completed the survey, while 74 percent of web-form respondents completed the survey after consenting, which is not statistically significantly different than the chatbot rate ($p = 0.611$).[9] In contrast, the completion rate

[9]This information was only available for those randomly assigned to the Qualtrics platform.

for phone surveys after consenting is 100 percent.

The shift to an outside website appears to be a barrier to web-form completion. Among web-form surveys, most dropouts (roughly 75 percent) occurred between clicking the ad, which took them to an external website, and inputting a first response. Among chatbot respondents, this source of attrition is less important, comprising less than 40 percent of dropouts. Instead, chatbot respondents are more likely to drop out during the consenting process, accounting for 22 percent of dropouts, versus fewer than 5 percent among web-form respondents. Drop-out timing is relatively evenly spaced through the survey modules among those who consent but do not complete the survey.

In terms of duration, the median completed phone survey lasted 43 minutes, while the median online survey lasted 26 minutes. Respondents who self-administered surveys by chatbot rather than by web form took longer on average, with a median of 18 minutes for web forms and 28 minutes for chatbots.

5.2 Respondent characteristics

Table 2 shows teacher demographic characteristics and technology use separately by round and survey modality alongside statistics generated from administrative data on the universe of public school teachers.[10] Columns 7–9 show unadjusted p-values from three types of t-tests: comparing all web form versus chatbot respondents, comparing all online to all phone respondents, and comparing all study participants to all teachers nationally.

Teachers in the Philippines are relatively young, with nearly 40 percent under age 35 and 91 percent age 55 or younger (Table 2, column 6). The age distribution among online survey participants is similar, albeit a little older, with 27–34 percent under age 35, and 91–94 percent ages 55 or younger. The phone survey sample, in fact, skews younger relative to the national population, with 50 percent under age 35, reflecting that older teachers were

[10]Technology use is based on a nationwide survey of all public school teachers conducted by DepEd in July 2020, with a response rate of approximately 98%. The other demographic characteristics are based on DepEd aggregated records shared with the research team.

less likely to share their information or agree to participate.

Overall, 82% of employed teachers are female (Table 2, column 6), while the share female is lower among both online and phone samples, averaging 71% and 76%, respectively in round 2. However, the gender distributions are not equal across online modalities. In rounds 1 and 2, chatbot respondents are 6 and 4 percentage points, respectively, more likely to be female than web-form respondents.

Technology usage is high among teachers nationally, as 87 percent owned a computer at home and 94 percent owned a smartphone or tablet as of May 2020. For this reason, the sample is well-suited to online remote survey methods, and the rates of smartphone ownership are comparable among chatbot respondents (89–91 percent), although it is considerably lower for those who responded via web survey (82–83 percent). It is notable also that smartphone ownership is not universal among online study participants, but nearly all teachers have access to a smartphone, tablet, or computer at home. Considering round 1 respondents (for whom I ask about ownership rather than usage for remote teaching), 89 percent of online respondents own a smartphone or tablet, 79 percent have a computer at home, and 3 percent have neither.

Device ownership does not necessarily imply internet access, as just 43 percent of teachers nationally had WiFi at home, and 63 percent had an active data plan on their smartphone or tablet. The rate of home WiFi access rates is relatively comparable to the share of teachers who report that they currently or plan to use home WiFi to reach students in both online surveys (36 percent) and phone surveys (43 percent). The national share of teachers who connect at home with data plans differs more from the online (50 percent) and phone survey (73 percent) responses. However, both measures should be interpreted with some caution, as comparisons also capture differences in question-wording between the survey and national data.

Reflecting the slightly older age profile among online survey respondents, they are also more experienced. One-fourth of online survey respondents have five years or less teaching

experience, while one-half have more than 10 years. Conversely, a bit more than one-third of phone survey respondents has five years or less teaching experience, and only 37 percent have more than 10. All public school teachers have completed at least a bachelor's degree, and among the more experienced set of online survey respondents, roughly half have completed a graduate degree, compared with one-fourth of phone respondents.

Another potential concern about online surveying is the difficulty of reaching teachers in remote areas. Figure 3 shows that both phone surveys with the DepEd-provided sample and online surveys from a social-media sample reached respondents from all regions in the Philippines, including those like the Bangsamoro Autonomous Region in Muslim Mindanao (BARMM), which is particularly remote. However, online recruitment overrepresents teachers in Metro Manila (NCR), who comprise 14 percent of online respondents but 8 percent of the total teacher population.[11]

5.3 Impacts of modality on response quality

5.3.1 Straightlining

I first examine how survey modality affects item differentiation, with a reduction in variation suggesting an increase in satisficing behavior, which increases survey noise and reduces statistical power (Oppenheimer, Meyvis and Davidenko, 2009). I measure "straightlining" by generating a binary variable for whether respondents use a single response category for all items in a battery, which I average across three batteries in this survey: the PHQ-4 for mental health, a 3-item set about the impacts of COVID-19, and a 7-item well-being inventory that measures distress (asked only in the second round). Columns 1 through 2 of Table 3 show the results.

The average straightlining rate is 24.3 percent for phone survey respondents.[12] While web forms have a 8.7 percentage point higher rate of straightlining, chatbots have a 0.9

[11] Appendix Table A.1 lists the region-specific respondents shares.

[12] The phone survey rates are 31 percent for mental health, 36 percent for COVID-19 impacts, and 5 percent for the well-begin inventory.

percentage point *lower* rate of straightlining relative to CATI, a difference of 10.9 percentage points between the two online modalitites. Differences between web forms and CATI and between web forms and the chatbot are statistically significant at the one-percent level, and weighting affects neither the magnitude nor statistical significance of the results.

To more flexibly capture item differentiation, I also measure each respondent's probability of differentiation, P_d using the scale point variation method (Linville, Salovey and Fischer, 1986; Krosnick and Alwin, 1988; McCarty and Shrum, 2000; Heerwegh and Loosveldt, 2008; Kim et al., 2019). I calculate P_d, where $P_d = 1 - \sum P_i^2$. P_i is the share of values rated at point i on the rating scale across the battery. I average P_d across all three batteries to generate an averaged differentiation index.[13] Results in columns 3 and 4 of Table 3 mirror the straightlining patterns in columns 1 and 2. Chatbot surveys substantially increase item differentiation relative to web-form surveys to the extent that the differentiation index is statistically indistinguishable between chatbot surveys and CATI (column 4).

One consideration is that some straightlined responses will reflect true preferences (i.e., a respondent experiences no symptoms of depression or anxiety), thus potentially confounding inattention with a willingness to disclose sensitive information. If that is the case, I would expect that lower rates of differentiation would be associated with lower rates of disclosure. In contrast, web-form surveys have lower rates of differentiation *and* higher rates of disclosure of sensitive topics. Additionally, Appendix Table A.2 reports impacts separately by question battery, and it shows similar differentiation patterns among potentially less-sensitive (COVID-19 experience) questions.

5.3.2 "Don't know" and refusals

In Table 4, I report the share of questions for which respondents answered "don't know", refused, or skipped a question.[14] While phone survey respondents answered that they didn't

[13] See Kim et al. (2019) for alternative measures of item non-differentiation.

[14] I exclude legitimate "don't know" answers, such as if a teacher did not yet know her school's remote learning plan.

know, refused, or skipped a question for an average of 0.7 percent of questions (column 1), this rate was four times higher among web-form respondents (2.4 percent), which is statistically significant at the one-percent level. Conversely, the likelihood of missing responses for chatbot respondents is not only 3.0 percentage points lower than for web respondents, but also it is lower than for phone respondents, which is statistically significant at the 1-percent level. As before, weights have minimal effect on the magnitude and precision of these estimates.

5.3.3 Survey duration

I next consider the impact of survey modality on duration, with results in Table 5. Shorter duration surveys are not inherently problematic; for example, more educated and younger respondents often complete surveys more quickly (Yan and Tourangeau, 2008). However, extremely short surveys most likely indicate inattention (Revilla and Ochoa, 2015). I define a low-duration survey as one that is less than 1.5 standard deviations below the mean duration overall (after trimming responses exceeding 120 minutes), which comprise surveys of 7 minutes or less. While the likelihood of low-duration surveys is higher for online survey respondents, the magnitude is small (0.3 percentage points) and not statistically significant (95 percent confidence interval: -0.005, 0.011), and there is no detectable difference for chatbot respondents when compared to web-form or CATI respondents.[15]

5.3.4 Potentially sensitive questions

Finally, I consider respondents' willingness to answer potentially sensitive questions. In columns 1–2 of Table 6, I show the impact of modality on the PHQ-4 index, for which higher values correspond to increased incidence of anxiety and depression, normalized based on CATI responses. Column 1 shows that PHQ-4 scores of web-form respondents are 0.28 stan-

[15]As Table 1 shows, chatbot surveys typically took longer than web-form surveys, likely reflecting that they necessarily show one question at a time. Neither the the web-form nor chatbot surveys did not have encoded pauses, although the time to load the next web-form page could have marginally increased survey duration.

dard deviations higher than CATI respondents, and chatbot respondents are 0.47 standard deviations higher (summing the two coefficients), statistically significant at the 1-percent level

Columns 3–4 of Table 6 shows similar results when asking respondents about their distress in seven domains of their life.[16] A higher number indicates more distress. I average these responses and standardize the mean and standard deviation of by phone-survey responses. Web-form respondents report 0.9 standard deviations more distress relative to CATI respondents. In the case of the distress index, chatbot respondents report 0.16 standard deviations less distress relative to the web-form respondents, although this is only statistically significant at the 10-percent level.

5.4 Selection into participation and response distributions

The differences in selection into participation (Table 2) and in response quality from the previous section are important because of their ultimate implications for the distribution of survey responses. Indeed, Table 7 shows that the distribution of responses varies substantially across online and remote modalities. The left panel of unweighted responses shows the distribution of responses across all surveyed teachers (column 1) and separately for web survey respondents, chatbot respondents, and phone respondents (columns 2–4). Columns 5 and 6 show p-values for a test of equality of means between the web and chatbot and between the phone and online modalities, respectively. I restrict the sample to round 2 respondents because teaching resources used and attitudes changed substantially once the remote school year begain.

The first two unweighted rows mirror results from Table 2, showing that rates of device ownership are comparatively lower for online respondents than phone respondents, although weighting removes this difference. The next six rows describe teachers' experiences with

[16]Specifically, I ask respondents how often they have been bothered by seven specific topics (workload, change in work environment, situation of students, relationship with colleagues, family concerns, finances, and national and community news). This inventory was asked only in the second round.

online teaching. Here, results do not differ between chatbot and web-form respondents, although phone respondents differ on several dimensions. They are much more likely to report that their schools provided with them a device like a laptop or tablet for remote teaching, they are slightly less likely to say they are confident in their ability to teach remotely, and they are more likely to say that they assess students at least weekly. Weighting on observable characteristics across both samples reduces these differences only slightly.

The third set of responses pertain to respondent attitudes, and here is where modality is associated with the largest differences, reflecting Table 6 on potentially sensitive questions. While 45–54 percent of teachers said that COVID-19 had impacted their finances a lot, only 38 percent of teachers surveyed by phone said the same. Similarly, while 45 percent of teachers reached online said they are now less able to maintain work-life balance, only 18 percent of teachers reached by phone said the same. Some responses are more comparable, such as whether COVID-19 had affected their lives (76 percent say it has "a lot") and time spent tending to personal needs (39 percent say it is less than before the pandemic).

Weighting the data to reflect national distributions of teachers on observable dimensions (by region, gender, grade level taught, and laptop ownership) does modestly reduce differences between phone and online survey respondents, but most differences from the unweighted data remain large and statistically significant. This highlights that the net effect of differences in data quality, along with differences in unobservable characteristics, affects survey findings across multiple dimensions.

The difference in sampling process between the CATI and online methods prevents a causal interpretation of the observed differences in response quality. Even after holding constant region and observable characteristics, other factors may predict both participation and survey quality. Conditional on the assumption of proportional selection, I calculate that selection on unobservables would have to be 1.5–2.4 times more important than the role of unobservables on the data quality measures in Tables 3 and 4 for unobservables to wholly account for the difference in quality between online surveys and CATI (Altonji, Elder and

Taber, 2005; Oster, 2019)[17], assuming that selection on unobservables is proportional to selection on observables. In the case of willingness to report on potentially sensitive topics (Table 6), selection on unobservables would have to be 0.8–1.1 times as important as the included observable factors.

6 Costs

The per-respondent marginal costs shown in Table 8 reflects the sum of modality-specific costs divided by the number of completed surveys. Because fixed costs such as staff time to draft, translate, and program survey instruments are unaffected by modality and are highly context-specific, I omit them to focus on those costs specific to each modality: namely advertising, survey platform, interviewer salaries, phone load, interviewer supplies, and respondent tokens.[18] On top of fixed costs, phone surveys add $6.28 per respondent, while web-form surveys average $0.92 per respondent and chatbots average $0.38 per respondent.[19]

While platform costs are likely to be relatively constant across contexts, Facebook advertising algorithms generate substantial variation in costs based on the characteristics of the targeted sample. For this study, online survey recruitment averaged $0.33 per completed survey, which is in line with Rosenzweig et al. (2020) who estimate a median Facebook recruitment cost per survey of $0.13 in Mexico and $1.15 in Kenya,[20] and Pham, Rampazzo and Rosenzweig (2019), who spend approximately $0.53 per completed survey in Kenya.

The primary costs of phone surveying are staffing ($6,697, or 77 percent of total costs) and respondent tokens ($1,320, or 15 percent of total costs), while online costs are limited

[17]For this exercise, I set R_{max}, to 1.3 times the measured R^2, following Oster (2019) and others, and I use the weighted estimates in the even-numbered columns. See results in Appendix Table A.8.

[18]Two potential, modest differences in omitted costs are the time to import the phone-survey sampling frame and any difference in software coding time between the chatbot platform and SurveyCTO platform used for the web and phone surveys.

[19]Tablets for surveying were borrowed from other projects, making total CATI costs a slight underestimate. The Qualtrics estimates in Table 8 omit the cost of a software license because these are usually held institutionally, and it would be impractical to purchase an annual license for a two-week survey of 157 respondents.

[20]Rosenzweig et al. (2020) note that costs increase when Facebook usage rates are lower and when engaging in more specific geographic targeting, as was the case in their Kenya sample.

to advertising and the survey platform.

7 Conclusion

This paper demonstrates that online surveys, and chatbots in particular, are an attractive means to conduct surveys cost-effectively in global settings with targeted populations. Using a randomized experiment of an online survey of K–12 teachers recruited through social media, I demonstrate that chatbot surveys yield much higher completion rates than web-form surveys (43 percent vs 29 percent) and produce higher quality data in terms of item differentiation and the frequency of "don't know" or skipped responses.

Comparing results with CATI surveys of teachers recruited through a pre-existing frame provided by the Philippine Department of Education, I find that web-form surveys are of lower quality, consistent with past research in high-income countries (Fricker et al., 2005; Heerwegh and Loosveldt, 2008). However, I also find that chatbot surveys perform equally well in terms of item differentiation and the likelihood of unrealistically short surveys, while they lead to fewer "don't know" or skipped responses relative to phone surveys. Additionally, both online measures lead to higher rates of disclosure on potentially sensitive topics. In practice, this leads to substantially higher rates of reported distress among teachers surveyed online rather than by phone. Web respondents report a 0.21-standard deviation higher PHQ-4 index relative to phone respondents, and chatbot respondents report a 0.57-standard deviation higher PHQ-4 index.

These results also show that social-media-based recruiting can yield wide demographic and geographic coverage in lower-income countries in which the targeted population has relatively high internet penetration rates. Although K–12 teachers have higher rates of smartphone and internet access than the general Philippine population (Schumacher and Kent, 2020), the continued rise in internet penetration and smartphone ownership (ITU, 2021), even just in recent years, indicates that the share of individuals reached by online

surveys and social media is likely to grow rapidly. For example, Facebook penetration rates—the share of active users divided by the population 15 and older, are over 80 percent across Latin America and Southeast Asia (Rosenzweig et al., 2021).[21]

One caveat is that recent evidence suggests that social-media recruitment may not effectively generate representative samples of the general population, typically overrepresenting younger people and men, in particular (Pham, Rampazzo and Rosenzweig, 2019; Rosenzweig et al., 2020). For those applications, the use of pre-existing sampling frames may provide better external validity.

However, a wide range of research questions requires narrower targeting, whether trying to reach members of particular occupations (as in this study), age groups, educations, religions, political affiliations, or to target along other dimensions. For these sorts of samples, social-media-based recruitment is likely to be particularly useful, as pre-existing sampling frames are less likely to be available for targeted samples, or they may be out of date. In the case of remote samples, high phone number turnover rates can render older frames less useful, and extensive targeting through random-digit dialing can be time-consuming and expensive.

Finally, I show that at less than $1 per respondent, the costs of implementing online surveys are a fraction of CATI, at $6.30 per respondent. While CATI methods are already less expensive than face-to-face surveying (Rosenzweig et al., 2021), additional cost reductions create additional opportunities for lower-cost and/or higher-frequency data collection, which is particularly important to expand the range of answerable economic questions and the pool of scholars who are able to contribute.

[21]Exact information on penetration rates is difficult to obtain, as the number of monthly active users may include duplicate and false accounts.(United States Security and Exchange Commission, 2019). In the Philippines, for example, the number of monthly active users exceeds the population ages 15 and older.

References

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of political economy*, 113(1): 151–184.

Amaral, Sofia, Lelys Dinarte, Patricio Dominguez, Steffanny Romero, and Santiago M. Perez-Vincent. 2022. "Talk or Text? Evaluating Response Rates by Remote Survey Method During COVID-19." *Evaluating Response Rates by Remote Survey Method During COVID-19*.

Arthi, Vellore, Kathleen Beegle, Joachim De Weerdt, and Amparo Palacios-Lopez. 2018. "Not Your Average Job: Measuring Farm Labor in Tanzania." *Journal of Development Economics*, 130: 160–172.

Assefa, Thomas W., Aditi Kadam, Nicholas Magnan, Ellen McCullough, and Tamara McGavock. 2022. "Who Is Asking and How? The Effects of Enumerator Gender and Survey Method in Measuring Intimate Partner Violence."

De Leeuw, Edith D. 2005. "To Mix or Not to Mix Data Collection Modes in Surveys." *Journal of Official Statistics*, 21(5): 233–255.

Dillon, Brian. 2012. "Using Mobile Phones to Collect Panel Data in Developing Countries." *Journal of International Development*, 24(4): 518–527.

Ellsberg, Mary, Lori Heise, Rodolfo Peña, Sonia Agurto, and Anna Winkvist. 2001. "Researching Domestic Violence Against Women: Methodological and Ethical Considerations." *Studies in Family Planning*, 32(1): 1–16.

Etang, Alvin, and Kristen Himelein. 2020. "Monitoring the Ebola Crisis Using Mobile Phone Surveys." In *Data Collection in Fragile States*. 15–31. Palgrave Macmillan, Cham.

Fafchamps, Marcel, and Bart Minten. 2012. "Impact of SMS-based Agricultural Information on Indian Farmers." *The World Bank Economic Review*, 26(3): 383–414.

Fricker, Scott, Mirta Galesic, Roger Tourangeau, and Ting Yan. 2005. "An Experimental Comparison of Web and Telephone Surveys." *Public Opinion Quarterly*, 69(3): 370–392.

Fulmer, Russell, Angela Joerin, Breanna Gentile, Lysanne Lakerink, and Michiel Rauws. 2018. "Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial." *JMIR Mental Health*, 5(4): e64.

Garlick, Robert, Kate Orkin, and Simon Quinn. 2020. "Call Me Maybe: Experimental Evidence on Frequency and Medium Effects in Microenterprise Surveys." *The World Bank Economic Review*, 34(2): 418–443.

Heerwegh, Dirk. 2009. "Mode Differences between Face-to-Face and Web Surveys: An Experimental Investigation of Data Quality and Social Desirability Effects." *International Journal of Public Opinion Research*, 21(1): 111–121.

Heerwegh, Dirk, and Geert Loosveldt. 2008. "Face-to-Face versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality." *Public Opinion Quarterly*, 72(5): 836–846.

Heise, Lori, and Mazeda Hossain. 2017. "Measuring Intimate Partner Violence." London School of Hygiene and Tropical Medicine.

Henderson, Savanna, and Michael Rosenbaum. 2020. "Remote Surveying in a Pandemic: Research Synthesis."

International Telecommunication Union. 2021. "Measuring Digital Development: Facts and Figures 2021." 31.

Jäger, Kai. 2017. "The Potential of Online Sampling for Studying Political Activists around the World and across Time." *Political Analysis*, 25(3): 329–343.

Kapp, Julie M., Colleen Peters, and Debra Parker Oliver. 2013. "Research Recruitment Using Facebook Advertising: Big Potential, Big Challenges." *Journal of Cancer Education*, 28(1): 134–137.

Kim, Soomin, Joonhwan Lee, and Gahgene Gweon. 2019. "Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality." *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Kim, Yujin, Jennifer Dykema, John Stevenson, Penny Black, and D. Paul Moberg. 2019. "Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail–Web Mixed-Mode Surveys." *Social Science Computer Review*, 37(2): 214–233.

Kreuter, Frauke, Stanley Presser, and Roger Tourangeau. 2008. "Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity." *Public Opinion Quarterly*, 72(5): 847–865.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology*, 5(3): 213–236.

Krosnick, Jon A., and Duane F. Alwin. 1988. "A Test of the Form-Resistant Correlation Hypothesis: Ratings, Rankings, and the Measurement of Values." *Public Opinion Quarterly*, 52(4): 526–538.

Krumpal, Ivar. 2013. "Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review." *Quality & Quantity*, 47(4): 2025–2047.

Lau, Charles Q., Eric Johnson, Ashley Amaya, Patricia LeBaron, and Herschel Sanders. 2018. "High Stakes, Low Resources: What Mode(s) Should Youth Employment Training Programs Use to Track Alumni? Evidence From South Africa." *Journal of International Development*, 30(7): 1166–1185.

Lee, Hana, Sunwoong Kim, Mick P. Couper, and Youngje Woo. 2019. "Experimental Comparison of PC Web, Smartphone Web, and Telephone Surveys in the New Technology Era." *Social Science Computer Review*, 37(2): 234–247.

Linville, Patricia W., Peter Salovey, and Gregory W. Fischer. 1986. "Stereotyping and Perceived Distributions of Social Characteristics: An Application to Ingroup-Outgroup Perception." In *Prejudice, Discrimination and Racism.* , ed. J. F. Dovidio and S. L. Gaertner, 165–208. Orlando, FL:Academic Press.

McCarty, John A., and Larry J. Shrum. 2000. "The Measurement of Personal Values in Survey Research: A Test of Alternative Rating Procedures." *Public Opinion Quarterly*, 64(3): 271–298.

Offer-Westort, Molly, Leah R. Rosenzweig, and Susan Athey. 2021. "Optimal Policies to Battle the Coronavirus "Infodemic" Among Social Media Users in Sub-Saharan Africa: Pre-Analysis Plan."

Oppenheimer, Daniel M., Tom Meyvis, and Nicolas Davidenko. 2009. "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power." *Journal of Experimental Social Psychology*, 45(4): 867–872.

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics*, 37(2): 187–204.

Park, David Sungho, Shilpa Aggarwal, Dahyeon Jeong, Naresh Kumar, Jonathan Robinson, and Alan Spearot. 2021. "Private but Misunderstood? Evidence on Measuring Intimate Partner Violence via Self-Interviewing in Rural Liberia and Malawi."

Pham, Katherine Hoffmann, Francesco Rampazzo, and Leah R. Rosenzweig. 2019. "Online Surveys and Digital Demography in the Developing World: Facebook Users in Kenya." *arXiv:1910.03448 [cs].*

Revilla, Melanie, and Carlos Ochoa. 2015. "What Are the Links in a Web Survey among Response Time, Quality, and Auto-Evaluation of the Efforts Done?" *Social Science Computer Review*, 33(1): 97–114.

Rick, Steven R., Aaron Paul Goldberg, and Nadir Weibel. 2019. "SleepBot: Encouraging Sleep Hygiene Using an Intelligent Chatbot." 107–108.

Rosenzweig, Leah, Parrish Bergquist, Katherine Hoffmann Pham, Francesco Rampazzo, and Matto Mildenberger. 2020. "Survey Sampling in the Global South Using Facebook Advertisements."

Rosenzweig, Leah R., Bence Bago, Adam J. Berinsky, and David G. Rand. 2021. "Happiness and Surprise Are Associated with Worse Truth Discernment of COVID-19 Headlines among Social Media Users in Nigeria." *Harvard Kennedy School Misinformation Review.*

Schumacher, Shannon, and Nicholas Kent. 2020. "8 Charts on Internet Use around the World as Countries Grapple with COVID-19."

United States Security and Exchange Commission. 2019. "Facebook Inc 2019 Annual Report 10-K." Washington, DC.

Winter, Nick. 2018. "SURVWGT: Stata Module to Create and Manipulate Survey Weights." *Boston College Department of Economics.*

Yan, Ting, and Roger Tourangeau. 2008. "Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times." *Applied*

Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 22(1): 51–68.

Figures

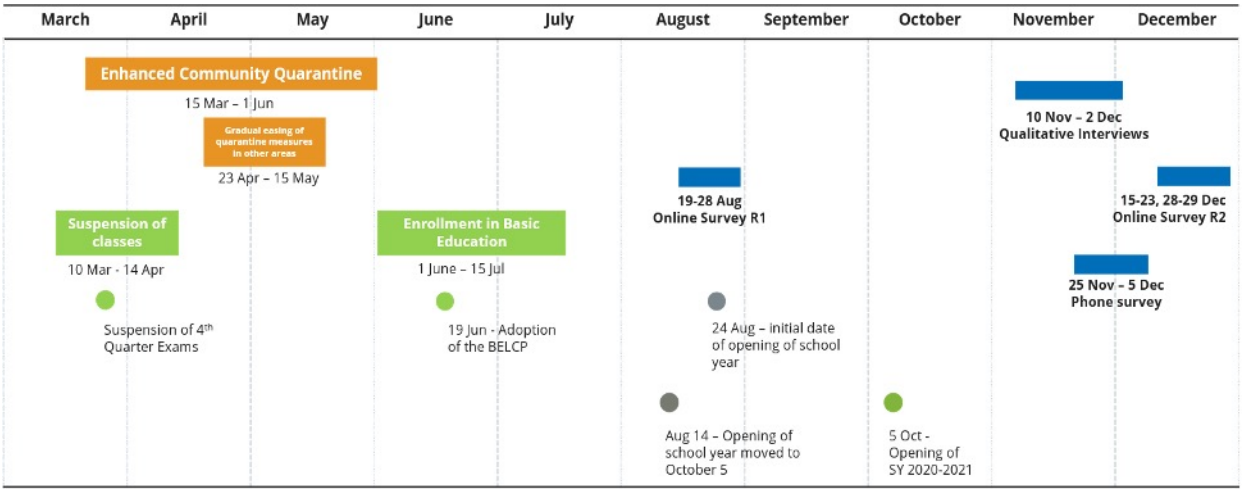


Figure 1: Study timeline

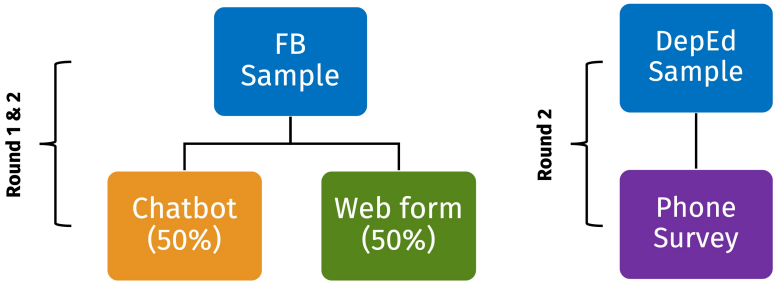
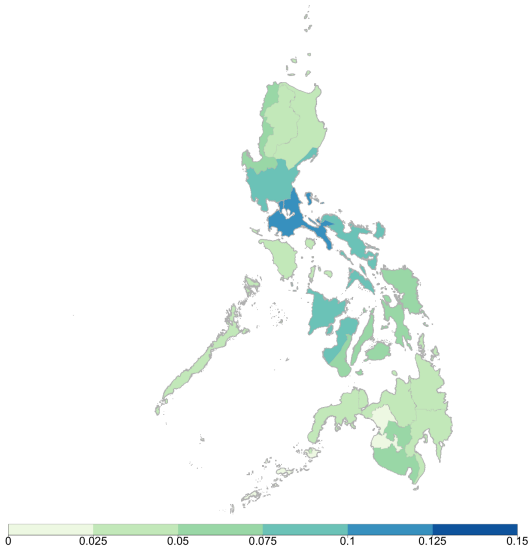
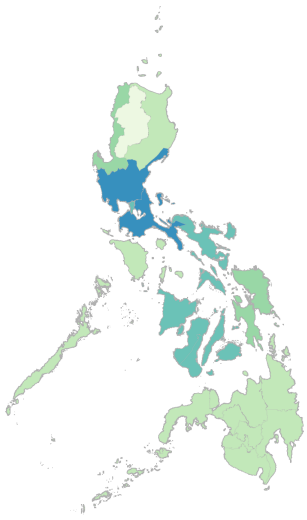


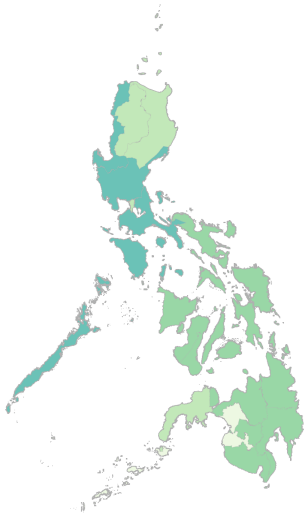
Figure 2: Randomization



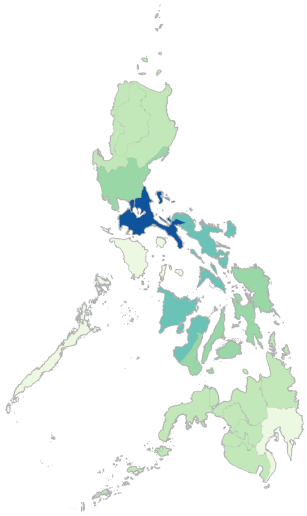
All respondents



K-12 teachers, nationally



Phone respondents



Online respondents

Figure 3: Geographic distribution of respondents

Tables

Table 1: Completion and response rates, by round

	Round 1		Round 2				Total	
	Web form, SurveyCTO	**Chatbot**	**Web form, SurveyCTO**	**Web form, Qualtrics**	**Chatbot**	**Phone**	**Web form**	**Chatbot**
Attempted						1,646		
Answered/started	1,461	2,264	537	534	1,285	1,367	2,532	3,549
Eligible		1,760		222	894	1,276		
Consented		1,537		212	716	1,229		2,253
Completed	415	1,163	169	157	541	1,229	741	1,704
Median duration (min.)	19.3	27.8	16.8	17.6	32.4	43.3	17.6	28.1
% Completed | Attempted						75%		
% Completed | Answered	28%	51%	31%	29%	42%	90%	29%	48%
% Completed | Consented		76%		74%	76%	100%		76%

SurveyCTO platform did not record incomplete surveys, so we lack full data on eligiblity and consent. Online surveys include 382 private school teachers excluded from analysis. Difference between those answered and eligible in phone survey reflect respondents who requested to reschedule but could not be reached in later attempts.

Table 2: Characteristics of respondents

	Round 1		Round 2			National	Unadjusted p-values		
	Web	Chatbot	*Web*	*Chatbot*	*Phone*		*Web vs. chatbot*	*Online vs. Phone*	*Sample vs. National*
	Currently own		*Use for remote teaching*			*Currently own*			
Computer at home	79%	79%	79%	79%	93%	87%	0.840	0.000	0.000
Smartphone/tablet	82%	91%	83%	89%	95%	94%	0.000	0.000	0.000
	Will use to reach students		*Use to reach students*			*Connect at home*			
WiFi at home	41%	29%	46%	42%	43%	43%	0.000	0.000	0.000
Data plan	39%	43%	54%	69%	73%	63%	0.042	0.000	0.000
Female	61%	67%	68%	72%	76%	82%	0.034	0.000	0.000
34 and below	33%	34%	27%	30%	50%	37%	0.301	0.000	0.005
35–44	37%	35%	34%	32%	30%	32%	0.559	0.021	0.196
44–55	23%	25%	31%	32%	15%	22%	0.894	0.000	0.554
56–64	6%	7%	9%	6%	4%	9%	0.296	0.003	0.000
<2 years experience	6%	7%	6%	7%	7%		0.428	0.502	
2–5 years experience	15%	20%	15%	18%	27%		0.015	0.000	
5–10 years experience	34%	26%	29%	23%	28%		0.004	0.483	
>10 years experience	45%	46%	51%	52%	37%		0.752	0.000	
Post-graduate[a]	46%	52%	57%	50%	26%		0.788	0.000	
Observations	335	968	263	451	1228	787,066[b]			

Notes: *a:* All K-12 public teachers have completed a bachelors degree, and "post-graduate" includes those who have completed a masters' degree or higher. *b:* Data on teacher technology use based on 787,066 teachers, while administrative records for age and gender include 798,151.

Table 3: Impact of modality on item differentiation

	Straightlining (0/1)		Differentiation index	
	(1)	(2)	(3)	(4)
Online	0.087***	0.114***	-0.038***	-0.050***
	[0.017]	[0.024]	[0.010]	[0.013]
Chatbot	-0.109***	-0.124***	0.058***	0.058***
	[0.016]	[0.023]	[0.009]	[0.012]
Observations	3259	3215	3255	3211
Mean, phone survey	0.243	0.243	0.371	0.371
p-value Online + Chatbot == 0	0.130	0.639	0.015	0.445
Weighted		X		X

Phone survey is omitted category. All specifications control for respondent gender, age, school type, position, respondent education, and whether has any children under 18 at home, along with day-of-week fixed effects. Missing covariate values flagged and recoded as zeros. Weights generated using iterative proportional fitting based on distribution of region X laptop ownership and gender X grade level among all public school K-12 teachers. Weighted regressions exclude individuals with missing values of variables used to generate weights.

Table 4: Impact of modality on missing responses

	Don't know/skip	
	(1)	(2)
Online	0.017***	0.026***
	[0.003]	[0.004]
Chatbot	-0.030***	-0.037***
	[0.004]	[0.006]
Observations	3292	3245
Mean, phone survey	0.007	0.006
p-value Online + Chatbot == 0	0.000	0.000
Weighted		X

Phone survey is omitted category. All specifications control for respondent gender, age, school type, position, respondent education, and whether has any children under 18 at home, along with day-of-week fixed effects. Missing covariate values flagged and recoded as zeros. Weights generated using iterative proportional fitting based on distribution of region X laptop ownership and gender X grade level among all public school K-12 teachers. Weighted regressions exclude individuals with missing values of variables used to generate weights.

Table 5: Impact of modality on survey duration (minutes)

	Duration < 1.5 sd	
	(1)	(2)
Online	0.003	0.004
	[0.004]	[0.003]
Chatbot	-0.007	-0.004
	[0.004]	[0.004]
Observations	3292	3245
Mean, phone survey	0.000	0.000
p-value Online + Chatbot == 0	0.279	0.998
Weighted		X

A -1.5 s.d. duration is 6.9 minutes or less. Phone survey is omitted category. All specifications control for respondent gender, age, school type, position, respondent education, and whether has any children under 18 at home, along with day-of-week fixed effects. Missing covariate values flagged and recoded as zeros. Weights generated using iterative proportional fitting based on distribution of region X laptop ownership and gender X grade level among all public school K-12 teachers. Weighted regressions exclude individuals with missing values of variables used to generate weights.

Table 6: Impact of modality on responses to potentially sensitive questions

	PHQ-4 (s.d.)		Distress index (s.d.)	
	(1)	(2)	(3)	(4)
Online	0.277***	0.210**	0.887***	0.757***
	[0.067]	[0.087]	[0.086]	[0.115]
Chatbot	0.195***	0.362***	-0.163*	-0.079
	[0.061]	[0.080]	[0.094]	[0.129]
Observations	3292	3245	1950	1932
Mean, phone survey	0.000	0.002	0.000	0.000
p-value Online + Chatbot == 0	0.000	0.000	0.000	0.000
Weighted		X		X

Phone survey is omitted category. All specifications control for respondent gender, age, school type, position, respondent education, and whether has any children under 18 at home, along with day-of-week fixed effects. Missing covariate values flagged and recoded as zeros. Weights generated using iterative proportional fitting based on distribution of region X laptop ownership and gender X grade level among all public school K-12 teachers. Weighted regressions exclude individuals with missing values of variables used to generate weights.

Table 7: Distribution of responses by modality

		Unweighted					Weighted					
	All	Web Survey	Chat-bot	Phone	P-value, Web = Chat-bot	P-value, Phone = Online	Web Survey	Chat-bot	Phone	P-value, Web = Chat-bot	P-value, Phone = Online	Obser-vations
Uses computer at home	0.88	0.79	0.79	0.93	0.926	0.000	0.88	0.87	0.87	0.827	1.000	1942
Uses smartphone/tablet	0.69	0.55	0.69	0.73	0.000	0.000	0.53	0.67	0.67	0.006	0.127	1942
School provided device	0.39	0.27	0.22	0.47	0.210	0.000	0.20	0.23	0.41	0.555	0.000	1931
School helps with out-of-pocket expenses	0.23	0.20	0.23	0.24	0.364	0.332	0.19	0.25	0.25	0.151	0.466	1930
Uses printed modular learning	0.98	0.96	0.98	0.99	0.166	0.001	0.99	0.98	1.00	0.714	0.030	1937
Uses online learning	0.39	0.38	0.43	0.38	0.212	0.178	0.32	0.38	0.38	0.252	0.535	1937
Confident in remote teaching	0.49	0.55	0.53	0.47	0.595	0.006	0.53	0.51	0.45	0.652	0.053	1932
Assess student performance at least weekly	0.33	0.24	0.23	0.39	0.684	0.000	0.27	0.23	0.37	0.369	0.000	1928
COVID-19 affected life a lot	0.76	0.75	0.72	0.77	0.370	0.035	0.71	0.68	0.79	0.504	0.001	1942
COVID-19 affected finances a lot	0.42	0.54	0.45	0.38	0.016	0.000	0.54	0.45	0.40	0.093	0.018	1942
Worried about someone in HH contracting COVID-19 a lot	0.76	0.75	0.70	0.78	0.126	0.002	0.75	0.70	0.77	0.288	0.073	1942
Relative to before the pandemic ...												
More hours working	0.68	0.68	0.63	0.70	0.136	0.023	0.73	0.64	0.69	0.066	0.578	1942
Fewer hours attending to personal needs	0.39	0.41	0.39	0.39	0.633	0.567	0.38	0.42	0.38	0.361	0.474	1942
Less able to balance work-life	0.27	0.44	0.45	0.18	0.783	0.000	0.42	0.42	0.18	0.964	0.000	1942
Less motivated about work	0.23	0.29	0.26	0.20	0.498	0.001	0.22	0.20	0.16	0.732	0.067	1942
PHQ-4 anxiety and depression, normalized	0.11	0.21	0.37	0.0	0.071	0.000	0.13	0.42	-0.04	0.014	0.000	1942
Distress index, normalized	0.26	0.81	0.66	0.0	0.115	0.000	0.68	0.57	-0.02	0.427	0.000	1932

Sample restricted to Round 2 survey respondents. Weights generated using iterative proportional fitting based on distribution of region X laptop ownership and gender X grade level among all public school K–12 teachers.

Table 8: Survey costs, by modality

	Round 1		Round 2				Average	
	Chatbot	**Online, SurveyCTO**	**Chatbot**	**Online, SurveyCTO**	**Online, Qualtrics**	**Phone**	**Chatbot**	**Online**
Advertising	$ 310	$ 111	$ 292	$ 91	$ 85	$ -	$ 602	$ 287
Survey platform	$ 59	$ 198	$ 58	$ 198	$ -	$ 198	$ 116	$ 396
Salaries						$ 6,697		
Communications						$ 351		
Tokens						$ 1,320		
Supplies						$ 182		
Respondents	1163	415	541	169	157	1392	1,904	741
Cost/respondent	**$ 0.32**	**$ 0.74**	**$ 0.65**	**$ 1.71**	**$ 0.54**	**$ 6.28**	**$ 0.38**	**$ 0.92**

Notes: R1 Online, R2 Online-SurveyCTO, and R2 Phone reflects one-month SurveyCTO pricing. Qualtrics platform pricing and tablet costs assumed to be sunk, so marginal cost is zero. Including tablet purchase costs raise in-person survey cost to $8.82.

A Appendix Figures and Tables

Appendix figures

Figure A.1: Sample advertisement

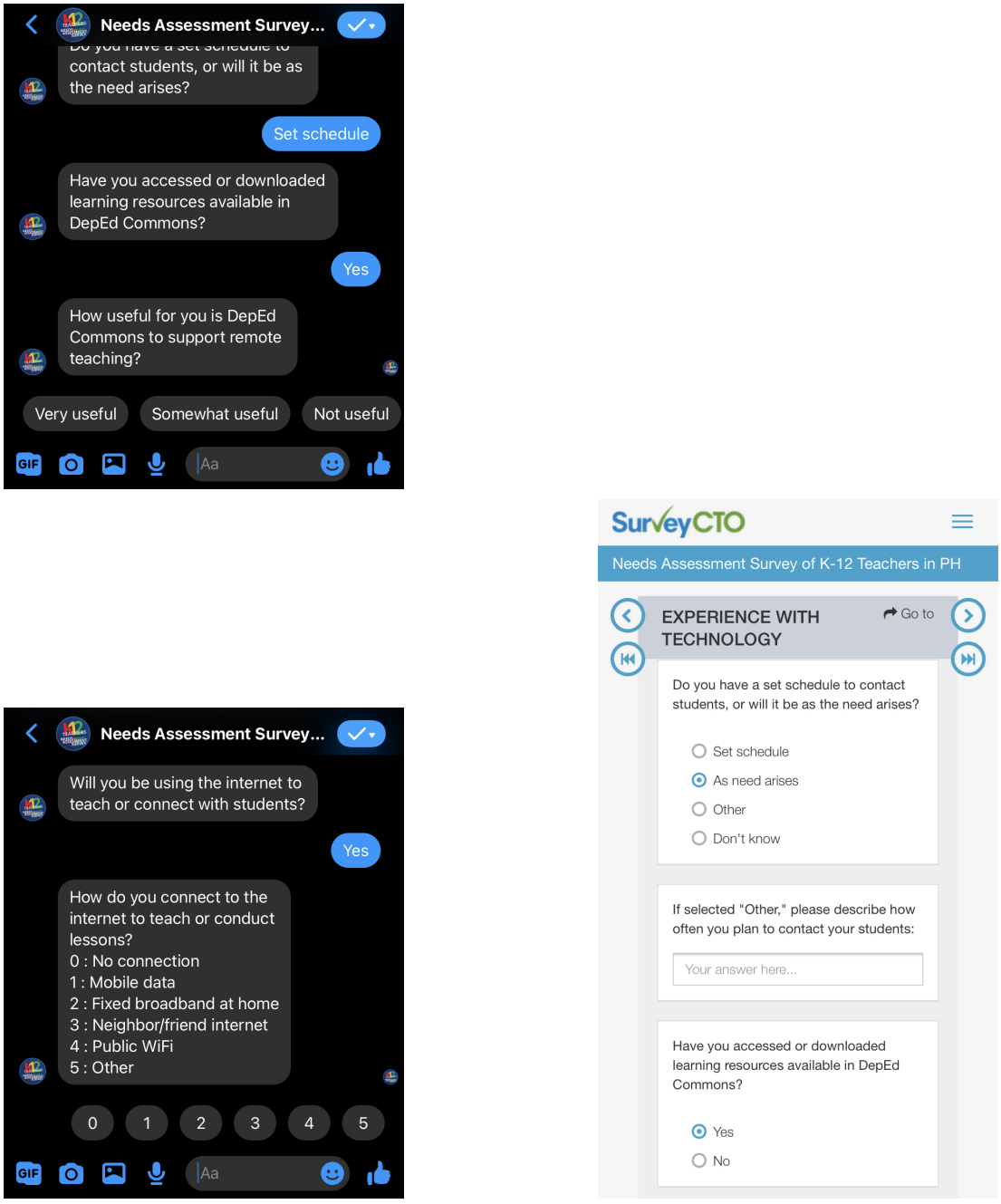


Figure A.2: User interfaces

Appendix tables

Table A.1: Distribution of response, by region

Region	Phone	Online	National
NCR	5%	14%	8%
CAR	4%	3%	2%
Region I	8%	5%	6%
Region II	4%	4%	4%
Region III	9%	7%	10%
Region IV-A	10%	13%	12%
Region IV-B	9%	2%	4%
Region V	5%	9%	8%
Region VI	7%	9%	9%
Region VII	5%	6%	8%
Region VIII	5%	6%	6%
Region IX	4%	4%	4%
Region X	5%	5%	5%
Region XI	5%	2%	5%
Region XII	7%	5%	5%
CARAGA	7%	3%	3%
BARMM	1%	3%	3%
Total	1,229	2,056	804,230

National data based on AY2020–2021 administrative records. Online sample excludes 7 observations with missing regional information

Table A.2: Impact of modality on straightlining, by battery

	Straightlining (0/1)				Differentiation index			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	PHQ-4	COVID-19	Distress	All	PHQ-4	COVID-19	Distress
Online	0.114***	0.032	0.148***	0.127***	-0.050***	-0.006	-0.070***	-0.037
	[0.024]	[0.041]	[0.042]	[0.037]	[0.013]	[0.022]	[0.020]	[0.024]
Chatbot	-0.124***	-0.130***	-0.102***	-0.127***	0.058***	0.063***	0.050***	0.020
	[0.023]	[0.034]	[0.036]	[0.040]	[0.012]	[0.018]	[0.017]	[0.026]
Observations	3215	3195	3204	1932	3211	3155	3186	1915
Mean, phone survey	0.243	0.316	0.359	0.055	0.371	0.326	0.299	0.487
p-value Online + Chatbot == 0	0.639	0.006	0.214	0.995	0.445	0.003	0.275	0.331
Weighted	X	X	X	X	X	X	X	X

Phone survey is omitted category. All specifications control for respondent gender, age, school type, position, respondent education, and whether has any children under 18 at home, along with day-of-week fixed effects. Missing covariate values flagged and recoded as zeros. Weights generated using iterative proportional fitting based on distribution of region X laptop ownership and gender X grade level among all public school K-12 teachers. Weighted regressions exclude individuals with missing values of variables used to generate weights.

Table A.3: Impact of modality on straightlining, alternative weighting

	Straightlining (0/1)				Differentiation index			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Online	0.087***	0.114***	0.091***	0.109***	-0.038***	-0.050***	-0.037***	-0.043***
	[0.017]	[0.024]	[0.020]	[0.024]	[0.010]	[0.013]	[0.010]	[0.013]
Chatbot	-0.109***	-0.124***	-0.103***	-0.130***	0.058***	0.058***	0.055***	0.063***
	[0.016]	[0.023]	[0.018]	[0.021]	[0.009]	[0.012]	[0.010]	[0.012]
Observations	3259	3215	3215	3215	3255	3211	3211	3211
Mean, phone survey	0.243	0.243	0.243	0.243	0.371	0.371	0.371	0.371
p-value Online + Chatbot == 0	0.130	0.639	0.444	0.330	0.015	0.445	0.048	0.094
Weighted		X	AW1	AW2		X	AW1	AW2

Phone survey is omitted category. All specifications control for respondent gender, age, school type, position, respondent education, and whether has any children under 18 at home, along with day-of-week fixed effects. Missing covariate values flagged and recoded as zeros. Weights generated using iterative proportional fitting based on distribution of region X laptop ownership and gender X grade level among all public school K-12 teachers. Weighted regressions exclude individuals with missing values of variables used to generate weights. AW1 generated based on distribution of region X laptop only, and AW2 generated based on distribution of region X gender X grade level.

Table A.4: Impact of modality on missing responses, alternative weighting

	Don't know/skip'			
	(1)	(2)	(3)	(4)
Online	0.017***	0.026***	0.017***	0.029***
	[0.003]	[0.004]	[0.003]	[0.007]
Chatbot	-0.030***	-0.037***	-0.029***	-0.042***
	[0.004]	[0.006]	[0.004]	[0.007]
Observations	3292	3245	3245	3245
Mean, phone survey	0.007	0.006	0.006	0.006
p-value Online + Chatbot == 0	0.000	0.000	0.000	0.001
Weighted		X	AW1	AW2

Phone survey is omitted category. All specifications control for respondent gender, age, school type, position, respondent education, and whether has any children under 18 at home, along with day-of-week fixed effects. Missing covariate values flagged and recoded as zeros. Weights generated using iterative proportional fitting based on distribution of region X laptop ownership and gender X grade level among all public school K-12 teachers. Weighted regressions exclude individuals with missing values of variables used to generate weights. AW1 generated based on distribution of region X laptop only, and AW2 generated based on distribution of region X gender X grade level.

Table A.5: Impact of modality on duration, alternative weighting

	Duration < 1.5 sd			
	(1)	(2)	(3)	(4)
Online	0.003	0.004	0.002	0.003
	[0.004]	[0.003]	[0.003]	[0.002]
Chatbot	-0.007	-0.004	-0.003	-0.003
	[0.004]	[0.004]	[0.004]	[0.002]
Observations	3292	3245	3245	3245
Mean, phone survey	0.000	0.000	0.000	0.000
p-value Online + Chatbot == 0	0.279	0.998	0.813	0.990
Weighted		X	AW1	AW2

A -1.5 s.d. duration is 6.9 minutes or less. Phone survey is omitted category. All specifications control for respondent gender, age, school type, position, respondent education, and whether has any children under 18 at home, along with day-of-week fixed effects. Missing covariate values flagged and recoded as zeros. Weights generated using iterative proportional fitting based on distribution of region X laptop ownership and gender X grade level among all public school K-12 teachers. Weighted regressions exclude individuals with missing values of variables used to generate weights. AW1 generated based on distribution of region X laptop only, and AW2 generated based on distribution of region X gender X grade level.

Table A.6: Impact of modality on responses to potentially sensitive questions, alternative weighting

	PHQ-4 (s.d.)				Distress index (s.d.)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Online	0.277***	0.210**	0.236***	0.205**	0.887***	0.757***	0.782***	0.932***
	[0.067]	[0.087]	[0.069]	[0.104]	[0.086]	[0.115]	[0.093]	[0.133]
Chatbot	0.195***	0.362***	0.253***	0.361***	-0.163*	-0.079	-0.088	-0.135
	[0.061]	[0.080]	[0.062]	[0.090]	[0.094]	[0.129]	[0.102]	[0.145]
Observations	3292	3245	3245	3245	1950	1932	1932	1932
Mean, phone survey	0.000	0.002	0.002	0.002	0.000	0.000	0.000	0.000
p-value Online + Chatbot == 0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Weighted		X	AW1	AW2		X	AW1	AW2

Phone survey is omitted category. All specifications control for respondent gender, age, school type, position, respondent education, and whether has any children under 18 at home, along with day-of-week fixed effects. Missing covariate values flagged and recoded as zeros. Weights generated using iterative proportional fitting based on distribution of region X laptop ownership and gender X grade level among all public school K-12 teachers. Weighted regressions exclude individuals with missing values of variables used to generate weights. AW1 generated based on distribution of region X laptop only, and AW2 generated based on distribution of region X gender X grade level.

Table A.7: Distribution of responses by modality, alternative weighting

	Region and laptop ownership					Region, gender, grade level					
	Web Survey	Chat-bot	Phone	P-value, Web = Chat-bot	P-value, Phone = Online	Web Survey	Chat-bot	Phone	P-value, Web = Chat-bot	P-value, Phone = Online	Observations
Uses computer at home	0.88	0.88	0.87	0.752	0.797	0.75	0.76	0.91	0.743	0.000	1942
Uses smartphone/tablet	0.55	0.71	0.71	0.000	0.016	0.54	0.64	0.67	0.045	0.073	1942
School provided device	0.26	0.22	0.46	0.271	0.000	0.21	0.23	0.44	0.631	0.000	1931
School helps with out-of-pocket expenses	0.21	0.24	0.23	0.403	0.965	0.17	0.22	0.25	0.164	0.056	1930
Uses printed modular learning	0.96	0.98	0.99	0.206	0.017	0.98	0.98	0.99	0.876	0.036	1937
Uses online learning	0.38	0.44	0.40	0.220	0.434	0.30	0.37	0.35	0.116	0.970	1937
Confident in remote teaching	0.55	0.52	0.47	0.547	0.017	0.51	0.52	0.46	0.849	0.044	1932
Assess student performance at least weekly	0.25	0.24	0.38	0.869	0.000	0.27	0.23	0.37	0.374	0.000	1928
COVID-19 affected life a lot	0.73	0.70	0.77	0.558	0.008	0.70	0.71	0.78	0.867	0.014	1942
COVID-19 affected finances a lot	0.53	0.45	0.39	0.041	0.001	0.52	0.44	0.41	0.141	0.077	1942
Worried about someone in HH contracting COVID-19 a lot	0.74	0.69	0.77	0.211	0.015	0.73	0.69	0.78	0.476	0.006	1942
More hours working relative to pre-pandemic	0.67	0.61	0.69	0.096	0.050	0.71	0.63	0.70	0.081	0.222	1942
Fewer hours attending to personal needs relative to pre-pandemic	0.39	0.40	0.41	0.705	0.562	0.40	0.42	0.35	0.631	0.053	1942
Less able to balance work-life relative to pre-pandemic	0.44	0.45	0.19	0.810	0.000	0.44	0.42	0.16	0.782	0.000	1942
Less motivated about work relative to pre-pandemic	0.27	0.25	0.20	0.562	0.016	0.25	0.21	0.16	0.374	0.010	1942
PHQ-4 anxiety and depression, normalized	0.17	0.36	-0.01	0.035	0.000	0.18	0.42	-0.03	0.061	0.000	1942
Distress index, normalized	0.72	0.62	0.01	0.340	0.000	0.79	0.61	-0.07	0.195	0.000	1932

Sample restricted to Round 2 survey respondents.

Table A.8: Relative importance of selection on unobserved factors

	R_{max}	δ
Straightlining	0.105	1.809
Differentiation index	0.097	1.547
Don't know/skip	0.418	2.371
PHQ-4 (s.d.)	0.096	0.800
Distress index (s.d.)	0.196	1.163

δ indicates the relative importance of selection on unobserved vs. observed factors, assuming that selection is proportional (Oster, 2019).