OPEN

DATA DESCRIPTOR

Social dynamics of short term variability in key measures of household and community wellbeing in Bangladesh

Md. Ehsanul Haque Tamal¹, Andrew R. Bell², Mary E. Killilea² & Patrick S. Ward³

Received: 17 December 2018

Accepted: 19 June 2019

Published online: 17 July 2019

High-frequency social data collection may facilitate improved recall, more inclusive reporting, and improved capture of intra-period variability. Although there are examples of small studies collecting particular variables at high frequency in the social science literature, to date there have been no significant efforts to collect a wide range of variables with high frequency. We have implemented the first such effort with a smartphone-based data collection approach, systematically varying the frequency of survey task and recall period, allowing the analysis of the relative merit of high-frequency data collection for different key variables in household surveys. This study of 480 farmers from northwestern Bangladesh over approximately one year of continuous data on key measures of household and community wellbeing could be particularly useful for the design and evaluation of development interventions and policies. While the data discussed here provide a snapshot of what is possible, we also highlight their strength for providing opportunities for interdisciplinary research in the household agricultural production, practices, seasonal hunger, etc., in a low-income agrarian society.

Background & Summary

Conventional household surveys typically use multiple visits with a large time gap between visits to construct longitudinal data¹. These data typically suffer from recall bias and lost intra-period variation, as respondents are asked in one sitting to recall events or outcomes that have transpired during the entire period between survey interviews^{2,3}. Furthermore, these surveys are very expensive to conduct⁴, with large enumeration teams visiting respondents, thus increasing the propensity for enumeration or data-entry error. The advent of Computer-assisted Personal Interviewing (CAPI) via laptop, tablet or smartphone, has made data collection cheaper, more reliable, and ultimately more efficient⁵. But even with CAPI tools at their disposal, these large household surveys require enumerators to engage with participants for an extended period of time to complete lengthy questionnaires, which can result in respondent fatigue, and consequently, poor quality and unreliable data³. It is also frequently difficult to collect data from remote areas and to conduct surveys during natural disasters, political disorder, etc., all of which are frequent in developing country contexts. The dataset described in this note provides some of the first evidence on the feasibility of an alternative to these costly, infrequent, difficult-to-conduct surveys. In particular, the data reported here were collected over the course of 50 weeks at relatively high frequency using smartphones. To ameliorate concerns of respondent fatigue, the long-form survey instrument (based on an integrated household survey instrument) was decomposed into small 'microtasks' that participants could address in the course of 5–10 minutes each. The study encouraged continued engagement and active participation in the data collection efforts through 'micropayments' that were awarded upon the successful submission of the microtasks, as well as credits toward the ownership of the smartphone. By putting the smartphones directly into the respondents' hands, our study could also capture data easily from those at the focus of study, even those in remote places where access is limited or unreliable.

This dataset contains approximately one year's worth of high frequency data containing information on a wide variety of experiences, inputs, and outcomes such as agricultural production and practices; experiences with

¹International Food Policy Research Institute, Dhaka, Bangladesh. ²New York University, New York, USA. ³Duke Kunshan University, Kunshan, China. Correspondence and requests for materials should be addressed to A.R.B. (email: andrew.reid.bell@nyu.edu)

Fig. 1 The multistage sampling scheme for participant selection.

income shocks and climatic events; household income, expenditure, consumption, labor, employment; migration; housing and sanitation; information and technology; as well as basic demographic and household characteristics.

The data collection was conducted with an Android based smartphone using a customized launcher for Open Data Kit (ODK) – an open-source, user-friendly, and easily deployable set of tools supporting data collection, visualization, and sharing, without the complications of setting up and maintaining one's own servers⁶. Survey were translated into Bangla and pretested on the application platform. The implementing partner, WIN Incorporated, assisted in the selection and training of participants, distribution of the smartphones, and provided oversight over the data collection process. A mobile operator, Banglalink, provided SIM cards, talk time, and mobile data to support the farmers as rewards based on their responses on the questions.

The data collection covered 480 farm households drawn from 40 villages in two subdistricts of Rangpur district in northwestern Bangladesh, selected using a multistage sampling technique. Farmers were selected randomly from a sampling frame that was prepared in consultation with the local agricultural extension officers, specifically focused on farmers' literacy and the likelihood that they would emerge as an early adopter of smartphone technologies. Participants responded regularly to 46 different survey tasks along a one-year period, with the frequency with which they received each task randomized to be either weekly, monthly, or seasonally (see Methods).

The design of this study allows measurement of recall bias and missed intra-period variation. Additionally, it allows identification of the most appropriate frequencies (weekly, monthly, or seasonally) to assess consumption, spending, labor, or other aspects of life experience⁷. Further, the dataset allows examination of quantities commonly collected as point estimates – such as subjective well-being – as time series and distributions. Particular examples of possible uses of the dataset include analysis of (i) intra-annual food security dynamics, with data on grain storage and 24-hour recall food diaries, (ii) consumption responses to exogenous shocks such as climate events or changes in the labor market, or (iii) linkages from access and use of clean water to illness and missed work or school. In sum, we expect the dataset to provide a high-frequency, intra-annually resolved window into the dynamics of topics commonly summarized as single numbers in an integrated household survey, albeit for a small and purposively selected sample.

Methods

Participant selection. Our sample was designed with a multistage sampling technique (Fig. 1) aimed at reaching farming households who were potential early adopters of smartphones. We selected Rangpur district in northwestern Bangladesh and based on the literacy rate reported in the Bangladesh National Census 2011, selected the two most literate upazillas (sub-districts) from the eight upazillas of Rangpur district, namely Mithapukur and Rangpur Sadar⁸. We then randomly selected 20 villages from each of these two upazillas, for a combined pool of 40 villages. Through our local implementing partner WIN Incorporated we assigned four local Points of Contact (POC) in each upazilla, who consulted with local agricultural extension officers to make a list (sampling frame) of technology early adopting farmers, who the officers deemed likely to be among the first to engage with smartphones. These farmer lists varied in length from 15 to 40 in different villages. We then randomly selected 8 to 16 farmers from these lists (proportionally depending on the length of the list) from different villages, for a total sample of 480 farmers. Over a series of half-day workshops in their home villages, these farmers were trained intensively to use the ODK application along with some other basic features of smartphone usage (see Survey Implementation).

Data collection platform. The data collection process was designed on ODK Collect for use on Google's open source Android operating system. It supports a form, survey, or algorithm into a sequence of input prompts that provide navigation logic, entry constraints, and repeating substructures⁶. The device distributed to the participants was the Symphony Roar V25 smartphone using Android 4.4.2, each of which were preloaded with the ODK application and a custom launcher "Data Exchange" developed by Nafundi (the developers of ODK; <https://nafundi.com/>). Participant responses were stored on ODK Aggregate server, a ready-to-deploy server that hosts forms and submitted results. It aggregates collected data and provides standard interfaces to extract data such as spreadsheets. ODK Aggregate is implemented on Google's App Engine and allows users to avoid the hassle of setting up one's own reliable service⁶. Across our project we used 20 Aggregate servers to capture data from 20 unique phone setups (see Smartphone Setups).

Smartphone setup. We structured our questionnaire as 46 short (5–10 minute) tasks, most of which were based on existing modules in the Bangladesh Integrated Household Survey⁹. Most of the survey items in these tasks are closed-ended with single or multiple selection options, while in a few cases participants were asked to input their answers in words or in numbers. Some questions asked them to take photos of places such as tubewells or latrines, or record a GPS location. While some tasks (for which relatively little intra-annual variation would be expected, such as information on farm plots or household members) were included only once in the experiment, most tasks were given to participants multiple times along the 50-week experiment at weekly, monthly, or seasonal frequencies. Additionally, many tasks included a 'crowdsourcing' component, in which the respondent would complete the task first for him- or herself, and then repeat the task for any friend, neighbor, or passerby; these participants were ideally selected on an approximately random basis, and their anonymity was ensured through the absence of personally identifying questions. The purpose of this crowdsourcing component in the study was to examine whether a crowdsourced sample might demonstrate a reduced selection bias, and thus better approximate a true representative sample, than our purposive sample of potential early adopters. Each crowdsourced task replicates the task completed by the sample respondent, with the additional request for basic demographic information (i.e., age, gender, literacy, education) at the beginning of the task. We do not assume that the same individuals are tracked over time in the crowdsourced sample. While this limits some uses of the crowdsourced sample, as it does not provide within-subject time series, it also means that the dataset includes a high-frequency cross-sectional sample that is not subject to panel conditioning (where repeated engagement with the same task affects the way a respondent performs it¹⁰). This inclusion possibly enables an identification of panel conditioning effects in our main sample, by providing an analogously high-frequency set of responses that are not completed repeatedly by the same respondents against which the main sample might be compared. The versions of each task included in the study are summarized in Online-only Table 1.

We constructed 20 unique smartphone setups (each given to 24 respondents, for a total of 480 participants in the sample). Each setup included exactly one version of each of the 46 tasks (e.g., Task 12, repeated monthly, not crowdsourced). We created these setups by first randomly assigning one version of each task (weekly, seasonally or monthly; crowdsourced or not) to each phone setup. Since versions with higher frequency (and those that included a crowdsourcing component) present both a higher respondent burden and a higher earning potential, and having the goal of standardizing both effort and earnings across setups, we then used a script (coded in Matlab) to make pairwise switches of task versions between smartphone setups until the Gini coefficient of earnings potential (from micropayments for participation in the survey throughout the duration of the project) across the phones fell below 0.001, indicating near perfect equality across setups¹¹.

Ethical approval. This study was reviewed and approved as a minimal risk application by International Food Policy Research Institute's (IFPRI's) Institutional Review Board (IRB #00007490, FWA #00005121). The IRB application number is 2015-49-EPTD-M, approved on 09/21/2015 under the title "Crowdsourcing Rural Data Collection via Android". All research staff working directly with this research were required to have completed IFPRI's CITI training course. Also, all field data collection staff were briefed in research ethics, including informed consent and data privacy. A written consent form (approved by the IRB) was signed by all the participants with one copy sent to IFPRI and one kept by the participants.

Survey implementation. Data collection was facilitated by WIN Incorporated, via eight Points of Contact (POC) from different villages of Rangpur. We conducted a Training of Trainers (ToT) with the POCs to provide a full orientation with the ODK system. The ToTs are a way to prepare new trainers with appropriate background, skills and hands on experience on the survey system to provide farmers proper knowledge and technical assistance. Following these trainings, the POCs conducted training with small groups of farmers (consisting of the 8–16 farmer participants from a particular village) to orient them on the full data collection approach and timeline. We used a custom ODK launcher application called "Data Exchange" in order to streamline participation for the participants (Fig. 2), tested along with several survey questions in a pretest with a separate sample of farmers to check the user friendliness of the platform, comprehensiveness of the questions and participant comprehension. Following the outcomes from the pretesting, we developed a training guidebook (in Bangla with visual instructions) adjusted survey item wording per the pretesting findings.

WIN Incorporated had multiple layers of team management in place to maintain the quality of the work. The eight POCs reported to and were guided by two local field managers based in Rangpur, who in turn reported to senior staff at WIN Incorporated and the project leads. The full process was monitored and scrutinized by IFPRI with frequent random checks. POCs were in close contact with the participants during the early stage of the study to troubleshoot any issues they faced. A common early issue was the loss of installed apps and tasks due to a manual resetting of the device. Consequently, POCs were trained to reinstall all project software to address this. Engagement with the experiment varied along the course of data collection, with average response rates to tasks

Fig. 2 Survey task experience of participants. **(a)** Participants are notified via a push notification that tasks are available. **(b)** When participants tap the notification or the app, the Data Exchange interface appears. **(c)** From which Android ODK is launched for the individual tasks – adapted from Bell *et al.*¹¹.

peaking at over 90% around the 10th week of the experiment and declining to around 40% at the 50th week of the experiment, with average response rates slightly higher for weekly tasks (Fig. 3).

Data Records

Project data are stored in the Harvard data repository¹² as standalone databases. The data can be categorized broadly into two types.

- (1) **Non-repeated modules:** For the most part, these are the demographic characteristics of the households, collected only once from the participants. There are three non-repeated modules: (a) Basic Information, (b) Household composition, and (c) Plots; b and c each contain nested data files that provide records for individual household members and plot level data, respectively.
- (2) **Repeated Modules:** These modules contain data on crops, production, consumption etc. A total of 97 files are available under this category including the nested data files. Typically, one repeated module has several nested data files, where nested files provide individual records for data types identified in the task (crops, animals, members, events, etc.); for example, the module “DrinkingWater” contains nested file named “DrinkingWater_source_level_2” to capture information about each of the different sources. The more complicated example “FertilizersAndPesticides” contains nested files:

Fig. 3 Response rates to weekly, monthly, and seasonal tasks along the length of the experiment.

- “FertilizersAndPesticides_plot_repeat_begin_level_2”,
- “FertilizersAndPesticides_crop_level_3”
- “FertilizersAndPesticides_group_crop_i_croptype_level_4”,
- “FertilizersAndPesticides_rep_fert_level_5”.

that include records for the plots treated, the kinds of crops treated in each plot, the specific crops within each crop type treated, and the specific treatments applied to these specific crops, respectively. The nested files are named with an underscore sign (“_”) and “level_#” after the main file names. A full list is presented in Online-only Table 2.

Technical Validation

We have maintained data quality across the following steps: (1) Questionnaire translation into Bangla adopting standard survey questionnaire, (2) questionnaire and ODK platform pretesting, (3) Recruiting tech-savvy POCs and farmers likely to be early adopters of smartphone technology, (4) ToT and thorough training on the ODK platform, 5) Data checking and cleaning from STATA and R.

The majority of the survey questions were adopted from the Bangladesh Integrated Household Survey (BIHS)⁹ and modified as required to fit the smartphone modality. BIHS survey questions and their translations into Bangla had previously been tested, and the use of exactly the same translation increases the acceptability of our survey questions. We further pre-tested the comprehensiveness of the questions and farmer comprehension after training in a session conducted with approximately 10 farmers in the village of Manikgong, close to Dhaka. We used four different modules whose items spanned the full range of input modes in the experiment, including single selection, multiple selection, taking photos, capturing GPS, etc. This pretesting enabled us to test both the questionnaires and the ODK platform from the users’ perspective. As our sampled participants were chosen from among likely early adopters of smartphone technology, we observed that it was generally easy for them to familiarize themselves with the full process of answering questions and the overall functionality of the hand-sets. Moreover, the POCs were chosen from among university students with good understanding of technology, facilitating troubleshooting with the participants after the culmination of the training. We conducted a ToT with the POCs and total of 40 half day training sessions (one in each village) with the participants. Each training was conducted by two POCs and coordinated by a field manager along with a senior staff from WIN Incorporated. We prepared a training manual translated into Bangla which the POCs used in training the farmers, and which was distributed to the farmers for their own reference. The trainings covered a brief overview of our data collection project, basic mobile care issues, basic and advanced functions of the smartphone, orientation with the data collection application (“Data Exchange”), an introduction to various sample survey forms, an outline of different survey tasks, and self-practice with several survey forms. Once the participants started submitting data, we checked their submissions on a daily basis from the ODK aggregate servers and sent them micropayments (mobile data and talk-time facilitated by Banglalink). Following completion of the near-one-year pilot study, data were collated and minimally cleaned prior to archival. We did not alter any data (e.g., censoring or other treatment of outliers) and did not discard any cases, as the use of logical checks and input bounding available within the ODK platform prevented the occurrence (to our knowledge) of unusable data. Our experiment did not include a ground-truthing component, and so we are not able to directly compare the quality of responses made via smartphone to those that might have been made using a traditional survey approach.

Usage Notes

Participants’ identification access. Publicly available data from this project redacts identifiers such as International Mobile Equipment Identity (IMEI) numbers, device numbers, phone numbers, etc., that could be used to identify participants, either directly or as the result of deduction. All GPS coordinates that are included in the dataset have been modified with minor random noise to prevent participant identification, though participants’ modified GPS locations will provide a proportionate spatial measure that allows the user to analyze geographical relationships. No other personally-identifying data were collected, and hence the confidentiality

and anonymity of the participants were fully ensured. Participants' name and location-related information was collected as part of their informed consent, but those data were not published anywhere and were kept locked in a secure location at IFPRI's Dhaka office with access only available to project researchers. We presume none of this will discourage any researcher to reuse the data. Any researcher requiring actual GPS coordinates, or any other personally-identifying information, is asked to contact the authors; sharing of any such data would require at a minimum ethical clearance from his/her institution, as well as signed data-use agreements that ensure the maintenance of respondent confidentiality and data security.

Potential for double counting. Modules tasking participants to collect information on tubewells and latrines may include records in separate responses identifying the same locations, as many respondents may share the same irrigation facilities and reside in the same villages.

Code Availability

The data are published as raw, without modification or recoding, as downloaded from the Google App Engine Servers, with the exception of GPS coordinates (see Usage Notes). Data from all unique phone setups and different frequencies were concatenated in single tables, and in some cases variables have been relabeled for clarity and consistency. The Matlab routines used to develop the initial smartphone setups and the summary information of the phone setups are included as Supplementary Files 1 and 2 to this publication.

References

1. Carletto, G. *et al.* Improving the Availability, Quality and Policy-Relevance of Agricultural Data: The Living Standards Measurement Study – Integrated Surveys on Agriculture, Third Wye City Group Global Conference on Agricultural and Rural Household Statistics (Food and Agricultural Organization of the United Nations, Rome). Available at http://www.fao.org/fileadmin/templates/ess/pages/rural/wye_city_group/2010/May/WYE_2010.2.1_Carletto.pdf. Accessed September 26, 2016 (2010).
2. Beegle, K., Carletto, C. & Himelein, K. Reliability of recall in agricultural data. *Journal of Development Economics* **98**, 34–41 (2012).
3. Deininger, K., Carletto, C., Savastano, S. & Muwonge, J. Can diaries help in improving agricultural production statistics? Evidence from Uganda. *Journal of Development Economics* **1**, 42–50 (2012).
4. Blumenstock, J. E. Fighting poverty with data. *Science* **353**, 753–754 (2016).
5. Leisher, C. A comparison of tablet-based and paper-based survey data collection in conservation projects. *Social Sciences* **3**, 264–271 (2014).
6. Anokwa, Y., Hartung, C., Brunette, W., Borriello, G. & Lerer, A. Open source data collection in the developing world. *Computer* **42**, 97–99 (2009).
7. Bell, A., Ward, P., Tamal, M. E. H. & Killilea, M. Assessing recall bias and measurement error in high-frequency social data collection for human-environment research. *Population and Environment*, **40**(3), 325–345 (2019).
8. Bangladesh Bureau of Statistics (BBS). *Bangladesh population and housing census 2011*. (Bangladesh Bureau of Statistics, Dhaka, 2011).
9. Ahmed, A. Bangladesh Integrated Household Survey (BIHS) 2011–2012” Datasets. *International Food Policy Research Institute*, (Washington, D. C. 2013).
10. Warren, J. R. & Halpern-Manners, A. Panel conditioning in longitudinal social science surveys. *Sociological Methods & Research* **41**, 491–534 (2012).
11. Bell, A. R., Ward, P. S., Killilea, M. E. & Tamal, M. E. H. Real-time social data collection in rural bangladesh via a ‘microtasks for micropayments’ platform on android smartphones. *PloS one* **11**, e0165924 (2016).
12. International Food Policy Research Institute (IFPRI), Department of Environmental Studies, New York University. Social Dynamics of Short-Term Variability in Key Measures of Household and Community Wellbeing in Rural Bangladesh. *Harvard Dataverse*, <https://doi.org/10.7910/DVN/HBQQVE> (2018).

Acknowledgements

We would like to thank Nafundi for their work in software development, and our partners Banglalink and WIN Incorporated for implementation in the field. We would also like to thank Matthieu Stigler for assisting with data processing. This work was supported by the Cereal Systems Initiative for South Asia (CSISA) of the Consultative Group on International Agricultural Research (CGIAR), with generous funding provided by the United States Agency for International Development (USAID) and the Bill and Melinda Gates Foundation. The authors also acknowledge support from the International Food Policy Research Institute and the CGIAR Collaborative Research Program on Policies, Institutions, and Markets.

Author Contributions

Conceptualization: A.R.B. and P.S.W. Questionnaire: A.R.B. and P.S.W. Software: A.R.B., P.S.W., M.E.K. and M.E.H.T. Project administration: M.E.H.T. Supervision: A.R.B. and P.S.W. Manuscript: M.E.H.T., A.R.B., M.E.K. and P.S.W.

Additional Information

Supplementary Information is available for this paper at <https://doi.org/10.1038/s41597-019-0128-0>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2019