

**Use of Digital Data Collection Tools in Large Scale Surveys:  
Experiences and Opportunities**

Padmini Sampath\*,

Ashwin Nagappa,

Arundhati Roy,

Ananya Chatterji,

Anusha Gajinkar,

Alpesh Gajbe

Archana Mehendale

# Contents

<b>1. Introduction</b> .....	4
<b>2. Context and the application</b> .....	5
<b>2.1. Pilot testing using the digital format</b> .....	6
<b>2.2. Results of pilot testing</b> .....	8
<b>3. Converting the tools on ODK format</b> .....	9
<b>4. Experiences and Findings</b> .....	15
<b>5. Concluding remarks</b> .....	18
<b>References</b> .....	18

## ABSTRACT

Technological advances in conducting research and its impact on data collection has been a subject of interest, particularly in the context of large scale surveys. Applications that can be used on mobile telephones or Personal Digital Assistants (PDAs) have transformed the manner in which data collection tools are designed, how data is gathered, how it is stored and made available for analysis (Couper, 2005; Zacharia, Lazaridou & Avraamidou, 2016; Ng, 2015). Although such tools have immense potential because of their ease of use and stability in terms of administration, literature on the use of such digital tools for data collection in developing countries, particularly the Indian context is sparse. Even though large scale survey is a commonly adopted research design among social scientists and development researchers in India, little is known about the use and experience of such digital tools for data collection.

Our paper attempts to fill this gap in literature by systematically outlining the process of using digital tools for data collection and discussing the considerations for tool design, testing, administration and data storage based on our first-hand experience of using the ODK Toolkit. We conducted a baseline as part of an impact evaluation study of Connected Learning Initiative (CLIX), a programme for improving quality of education in rural government high schools using technology. We administered a set of 11 survey instruments capturing textual, visual and audio data. The experience of developing the instruments in four languages - English, Hindi, Telugu and Mizo and using them in four states viz. Chhattisgarh, Mizoram, Rajasthan and Telangana is presented. Based on our experience of the pilot phase, we compare the experience of using digital tools with a traditional mode of administering paper-based forms. Lastly, the challenges and opportunities of using the ODK Toolkit with a wide range of respondents - high school students, teachers, officials - many of whom were digitally challenged, are critically discussed. Our main conclusion is that digital tools such as the ODK Toolkit offer several advantages over the traditional paper-based tools, especially in the case of collecting structured data questions. They also help to reduce the incidence of missing data, illegible entries, errors on account of data entry and the time gap between data collection and rendering data for analysis. However, we find the tool inappropriate for use in the case of unstructured question items.

**Keywords:** *Survey research, Data collection technology, Open Data Kit*

## **1. Introduction**

Technological advances in research and its impact on data collection has been a subject of interest, particularly in the context of large scale surveys. Applications that can be used on mobile telephones or Personal Digital Assistants (PDAs) have transformed the manner in which research data collection tools are designed, how data is gathered, how it is stored and made available for analysis (Couper, 2005; Zacharia, Lazaridou & Avraamidou, 2016; Ng, 2015). Jeffrey-Cocker, Basinger and Modi (2010) have discussed the use of Open Data Kit (ODK), a

free and open source tool for data collection with farmers in rural Mali. They assert that ODK Toolkit has a ‘profound impact on the future of data gathering’ (pp.1) because it can save time, costs and can be administered in remote locations with the help of Android based mobile phones. Brunette et al. (2013) have discussed how the developers’ team at University of Washington, Seattle, have refined the ODK Toolkit based on user feedback so as to expand the range of applications, and make it customizable by users.

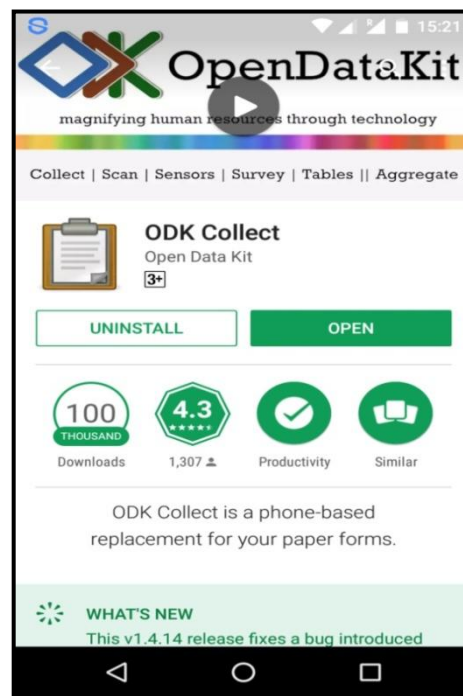
Literature on the use of such digital tools for data collection in the Indian context is sparse, although such tools have immense potential given their ease of use and stability in terms of administration. Even though large scale survey is a commonly adopted research design among social scientists and development researchers in India, little is known about the use and experience of such digital tools for data collection. Our paper attempts to fill this gap in literature by systematically outlining the process of using this particular application and discussing the considerations for tool design, testing, administration and data storage based on our first-hand experience of using the ODK Toolkit.

In this paper, we share some of our findings of using the Open Data Kit (ODK) in large scale surveys and highlight the potential and limitation of its application. Our lessons are generally applicable to large scale surveys conducted using similar research design, and not necessarily limited to Indian context where our experience is located.

## **2. Context and the application**

We conducted a baseline as part of an impact evaluation study of Connected Learning Initiative (CLIX)<sup>1</sup>, a programme for improving the quality of education in rural government high schools using technological affordances. We developed and administered a set of 11 survey tools meant to capture textual, image and audio data. The tools were developed in four languages - English, Hindi, Telugu and Mizo and administered in rural areas of four Indian states viz. Chhattisgarh, Mizoram, Rajasthan and Telangana. There were over 1000 high school students, 600 teachers,

200 principals and education officials as respondents. The decision to use Open Data Kit (ODK) was made because it is an open-source Android based application (see Figure 1), it is easy to use and helps save considerable amount of time and errors. Unlike other digitised data collection applications, the ODK Toolkit does not require continuous internet connectivity but requires it only at the time of downloading a blank form and at the time of uploading all the forms that are saved on the device. Moreover, we were particular that data collected through this application is stored on our own server instead of a third party remote location.



*Figure 1*

## **2.1. Pilot testing using the digital format**

The tools were first developed and pilot tested on the ODK format in four states in four languages. This was done in order to test if the data collection experience was more efficient using the digital tools compared to the conventional paper-pen based tools. This was also important because we wanted to confirm if the digital tools actually worked without continuous internet connectivity, whether the application was robust to handle multiple forms, multiple languages and three different kinds of data – textual, image and audio. But most importantly, we wanted to know if the tools would work in rural India where most of the respondents were not

digitally literate but were exposed to 'smart phones' and the associated interface. The long term motivation behind this exercise was to discover new ways of data collection and data processing, especially for large scale surveys.

The development of the tools for pilot testing involved populating a spreadsheet with the content from the data collection instruments in a predefined format, using simple macro like formulae. We found that there were chances of erroneous entries, and the chances of debugging long tools were becoming challenging. To overcome such challenges during the pilot testing phase, we tried using the Kobo Tool box, an enhanced version of ODK. Kobo made the process of digitisation easy given the intuitive Graphical User Interface (GUI) and ability to export the XML directly (as compared to converting the xls into xml). Although Kobo could also be used to collect data, the online form required continuous connectivity. The offline feature was not found to be stable to gather data using online form. Given the similarity of the apps in terms of their other features, it was decided to use the ODK instead of Kobo to collect data.

In order to test if a digital approach to data collection would work and its comparative advantages/disadvantages over the conventional paper-pen tools, we decided to administer the data collection instruments in both formats, where in half of our respondents would be given the digital tools on ODK and the rest would be given the same questions in a paper based format. Across the three states of Mizoram, Rajasthan and Telangana where we did the pilot testing, we had planned to administer the ODK tools to 60 students and 45 teachers and an equal number would be given a paper version. However, in Mizoram, all the teachers wanted to take the survey only in the digital form because they found it more interesting than the conventional mode. Thus, 60 teachers participated in the pilot testing with the ODK based tool and 30 teachers taking it in paper pen format. Since the number of students was small, we tried administering the tool on PDAs as well as on laptops with pre-installed Android emulator in schools where the computer labs had standalone computers as in the case of Mizoram. The computer labs used in Rajasthan and Telangana use thin-client systems and therefore we could not use the computers for administering ODK to a group of students at the same time.

## **2.2. Results of pilot testing**

The findings during the pilot testing were positive and encouraging. We found that although none of the respondents (students and teachers) had any prior experience of using any digital data collection tool and while a large number of them were also not digitally literate (hence the need for an intervention like Connected Learning Initiative), they were able to understand the operation of the ODK tool with little or no help. This could be largely on account of their exposure to Android based 'smart phones' that have a huge penetration even in rural India. There was also a sense of novelty and interest among respondents using the digital tools in comparison to those taking it in paper-pen formats. We also found that using a digital tool like ODK helped us to cut down the tasks related to data entry which was required in the case of paper-pen tools. The number of no-responses to questions on account of respondents not noticing the question or skipping them knowingly was much higher in the case of paper-pen tools compared to the ODK tool.

Our experience of testing the ODK tool with groups of students in their school computer lab was revealing. Firstly, we found that the configuration of computer systems available in the school labs in each of the states was different. In Mizoram, the Android emulator on standalone systems worked well but the time taken per student was more compared to the paper based format, because there was no 'swiping' function and students had to use the mouse or arrow keys to move to the next question. We saw shades of 'respondent fatigue' in response to the cumbersome process of answering the questions. In the states of Rajasthan and Telangana, where thin-client systems are used in the computer labs, we found that the process was even more time consuming because we could accommodate only one student at a time on the host computer. Thus, we found that administering the ODK tool simultaneously to a group of students in the school computer lab with thin-client systems as in the case of Rajasthan and Telangana would not work. Hence we decided to use the ODK based tools only where the survey instrument required a one-to-one interview/data collection method. If administered simultaneously to a group of respondents, we would require a set of PDAs or standalone computers with Android emulators.

Given the overall positive experience of the pilot testing, we decided to use the ODK Toolkit in the actual baseline survey. In the case of teachers, principals and officials who were to be interviewed, the ODK based tool was administered by the field investigators on a PDA/tablet. In

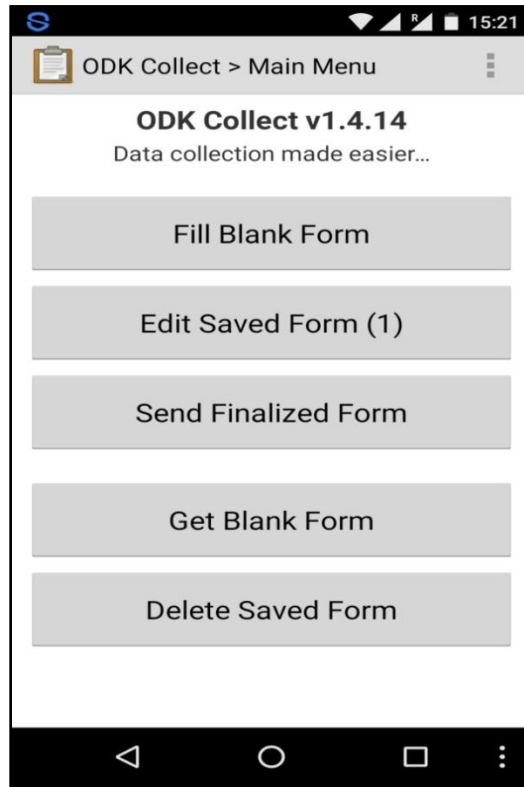


most cases, the respondents self-administered the tool, but in cases where they were not comfortable handling it on their own, our investigators helped to fill in the responses while showing the respondents the inputs being fed into the form. In the case of students, we decided to use the ODK tool only in Mizoram with some slight modifications on the interface. In the other states, we decided to use the conventional paper based forms.

### **3. Converting the tools on ODK format**

In this section, we explain the step-by-step process of creating the survey tools using ODK Toolkit. Although some of these instructions are available on the Help and Resources webpages of the official website of ODK, we are of the view that our actual experience of using the application in the field to collect large data would be valuable to researchers conducting surveys in developing countries. The question items in our eleven data collection instruments were textual, image and audio format. The expected responses were also in textual, image and audio formats. We think our experience of using multi-format data questions and being able to record multi-format data responses opens wide possibilities for researchers doing large scale surveys wanting to try interesting ways of data collection. We now share some of the learnings about the process of digitisation, how does it work and what is required to set it up.

The key requirements for using ODK in surveys are [1] a server where the blank survey forms can reside, which can have protected access through user logins, from where blank forms can be downloaded for purposes of administration and where the filled in forms can get uploaded and stored, (see figure 2 for a screenshot of the ODK Collect) [2] an Android device (either phone, tablet or computers with Android emulator) is required to collect data.

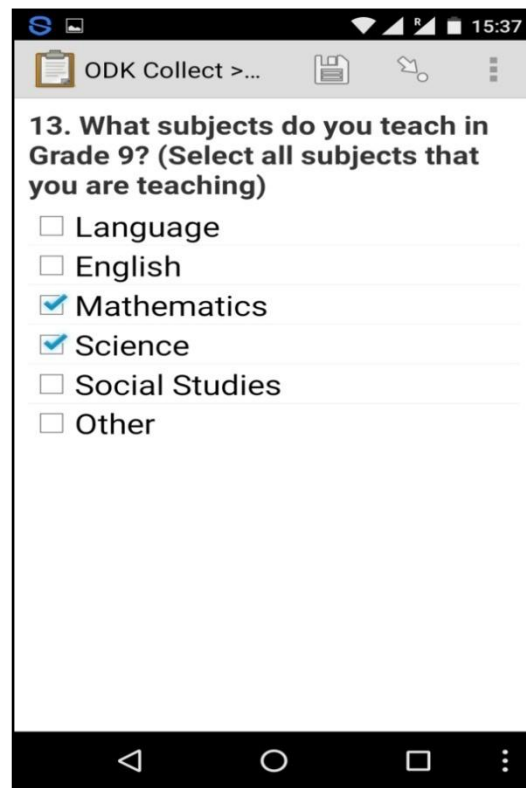


*Figure 2*

For building the forms on ODK Toolkit, the form is first authored and coded in a spreadsheet with a simple interface of survey (questions with the required structure), choices (specify answer choices) and settings (which is optional and can be used for overall structure and style). One of the features that we found useful was the provision to define questions as mandatory so that respondents are required to fill in that data, thereby reducing the incidence of missing data.

It is important to use pre-coded forms, so that the output generated in the form of spreadsheets contains coded responses rendering it immediately for analysis. We found that the ODK Toolkit is most suitable for administering structured tools containing different kinds of questions viz. single choice, multiple choice, ranking questions and questions using Likert scale (see figure 3 for an example of a multiple response question display). The feature of making questions conditional allowed us to reduce possibility of errors and improve internal validity. We found that ‘groups’ containing more than one question could appear on one screen under the common

heading, but visibility is compromised if the group has more than five questions. We could upload the images, collect the audio files as well as replay pre-recorded audio files.

A screenshot of a mobile application interface for ODK Collect. The top status bar shows the time as 15:37 and various system icons. Below the status bar, the app title "ODK Collect >..." is visible. The main content area displays a question: "13. What subjects do you teach in Grade 9? (Select all subjects that you are teaching)". Below the question, there is a list of subjects with checkboxes: Language, English, Mathematics, Science, Social Studies, and Other. The checkboxes for Mathematics and Science are checked. The bottom of the screen shows the standard Android navigation bar with back, home, and recent apps buttons.

*Figure 3*

Considerable time was required in ensuring that the questions are worded (content) in a manner that is appropriate to the method of rendering (form). Thus, questions needed to be clear, well-thought especially with regards to kind of responses expected, conditionalities, and how they would appear on the screen of the device. The tools once developed resided on the ‘test server’ from where we downloaded, validated, filled in trial forms, and uploaded them in order to test the entire loop. Tools in vernacular languages had to be coded in Unicode fonts so as to avoid errors in display of font (see figure 4.1, 4.2 and 4.3 for question displays in three languages). Once tested, the forms were uploaded on a ‘production server’ from where they were ready to be downloaded by the field investigators in the four states using their login access credentials. A database created on the ‘production server’ stored the forms and submissions which was tracked on the back end. The data was rendered in easily usable spreadsheet formats.

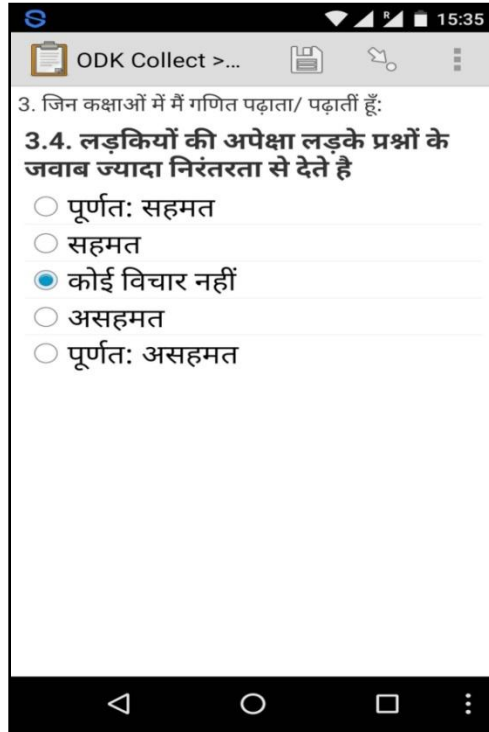


Figure 4.1

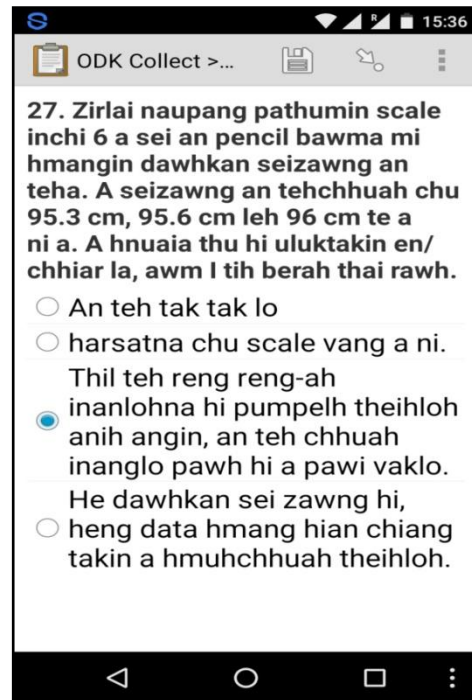


Figure 4.2

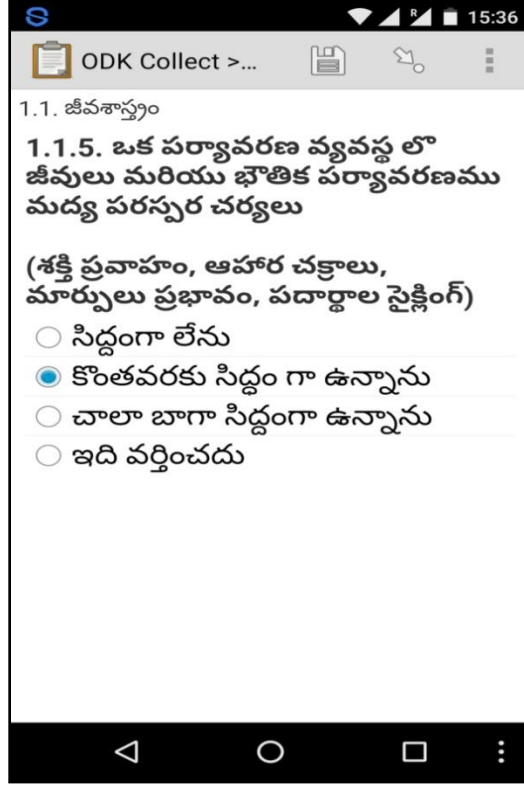
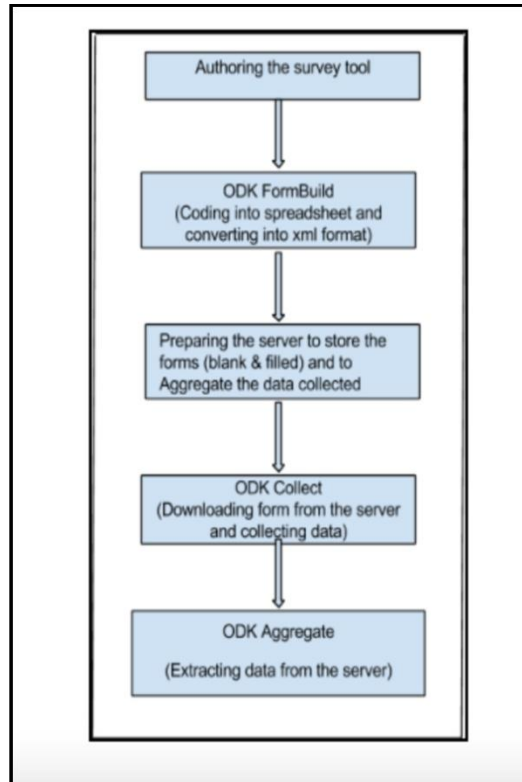


Figure 4.3

Form Build is the process of creating the forms in the ODK format. This requires a survey data collection instrument with all the questions in the required format and order. As first-timers to this process, we found the product website informative and instructive about creating new forms and aggregating data. For building the forms on ODK, the form is first authored and coded in a spreadsheet with a simple interface of survey sheet, choices sheet and settings sheet. In Survey worksheets we place the content with questions or heading in each row. There are three columns in the survey sheet viz. type column, name column, and label column. Type column is where we specify the nature of the answer (like integer, text, or as single choice or multiple choice, etc.) The name column is used to give variable name for each row. The label column has the questions and headings that are displayed to the end-user. Since we were translating the tools, we needed to add another label called translation column. We also had options of adding columns on survey sheet such as hint, constraint, relevant, required and few more. The choices worksheet is used to put answer options for multiple choice question type, these answer choices can be reused. The

choices sheet also has columns list\_name where we specified name of our choice, name column which is used in the survey sheet to link to these answer choices and label column where answer options are put which are displayed in the form. In the Settings worksheet which is optional, we added columns form\_title, which shows the title of form displayed to users; form\_id, where we specified the name used to identify the form.

One of the useful features of ODK Toolkit was the option of making responses to certain questions mandatory. As mentioned earlier, it helped to significantly reduce the incidence of missing data. Another feature to improve internal validity of the tool was the feature of making questions conditional to the kind of responses received elsewhere in the tool. For instance, for the question ‘do you read newspapers?’ if the teacher had answered ‘no’, the next question ‘what kind of news do you like to read?’ would not get displayed. In a conventional paper-pen self-administered format, a respondent could choose ‘no’ as the response to the first question and still fill in the news they liked to read, thereby raising questions on the validity and reliability of the response. A common practice in such cases in the traditional method is to discard both the questions resulting in loss of data. In the ODK tools, this can be controlled by inserting conditionality of specific response. For better readability, some of the questions were grouped together so that they appeared under a single heading. We could upload the images as well as audio files as part of the questions. We could also receive images and audio files as part of the filled in form. We had two servers, a test server, where we checked the forms before the field team could use and other is was called a production server where the final tool was available for use and also to hold the data uploaded from the field. A database was created on the production server so that all the forms get stored and updated as and when the submission of forms is done. A select number of team members also had access to the back-end of the production server where we could view the data and track the progress of submissions at the field level. Figure 5 summarises the overall workflow using ODK Toolkit from preparing the tool to extraction of data for analysis.



*Figure 5*

#### **4. Experiences and Findings**

As part of our baseline study of Connected Learning Initiative, we prepared and administered eleven ODK based digitised survey tools in four languages and used it across four widely diverse states in India for collecting data. We used these tools in over 200 schools, with students, teachers, principals and officials. We also used it to conduct over 450 classroom observations. On the basis of this rich experience, we conclude that a digitised tool such as ODK helps in easing the data collection process as well as significantly reduces the time taken to render the data for analysis after data gathering. Some of the key advantages and limitations are indicated below.

## **Advantages**

1] Considerable resources are saved because there is no need for printing, packaging, carrying printed tools, storing them, or for doing data entry. This allows the ODK Toolkit to be used widely in large scale surveys where these costs typically run very high.

2] Since the tools can be pre-coded, as soon as the filled in forms are uploaded, they became immediately available for analysis. There was no delay or error on account of data entry or post-coding the tools.

3] Internet was not required for collecting data because once the blank form is downloaded and saved on the device it could be used to gather data from multiple respondents, every time filling in a new form as a new instance. The forms are saved on the device itself until they are ready to be uploaded. Internet connectivity was required only to build the form, download a blank form and submit a filled up form. This made it user-friendly even in remote areas where internet connectivity was intermittent and feeble.

4] The interface on the application is simple and easy to use. Thus, it was possible to train field investigators who had no prior skills of using digitised survey tools. There were no issues/difficulties reported from the field with regards to administering the tool. Most of the respondents also wanted to self-administer the survey forms and we found that lack of digital literacy did not affect the ease with which the respondents were able to fill in the forms.

5] We were able to collect a range of textual, audio, visual data from the field since the ODK Toolkit has these capabilities. This made it easier to manage multiple kinds of survey tools on one application.

6] During the course of data collection, we were able to easily make slight modifications to the tool in the form of correcting language used in the questions. However, the corrected version of the tool had only to be uploaded again and the field investigators had to download the new forms and use them instead of the older version. Such mid-course rectification could be carried out without additional overheads unlike conventional tools that are extremely difficult to recall once they are being used. This also helped us to manage the entire data collection cycle from pilot testing of tools to actual data collection.



7] We also found that the respondents could also change their response to a question or even review their fully filled in form before finally submitting the same. Such an option is not available in the conventional paper-pen mode where the respondent may have already selected/checked a particular option. Such flexibility is helpful in reducing any anxieties that the respondents may have.

8] Barring a few cases (less than five in over couple of thousands) of forms that were presumably uploaded but were not found on the back-end server, there was no loss of forms that were filled and uploaded. Being in a digital form also meant that there was little or no scope for tampering/losing/damaging data once it is uploaded on server. We were also able to do a real-time (almost) check on data collected and received at the back-end allowing us to track the forms as they were being filled in and getting uploaded.

9] Even if a server is not connected, the filled in forms can be collected in a ODK Briefcase (offline) and shared with a central data co-ordinator. Although we did not have to use this feature, we had tested this as an option in case of problems connecting to the server for uploading data

.

## **Limitations/Challenges**

Some of the limitations we found were:

1] The ODK Toolkit was not suitable for collecting open-ended and qualitative data because as a default, it can store text data/strings up to 255 characters. We could increase this limit to a maximum of about 16000 UTF-8 characters. But we found that whenever the text data entered was more than 255 characters, it truncated the data to this limit without notifying the user and accepted the submission.

2] The interface was very challenging on an Android emulator installed computer machines as the swiping feature was not available and the respondents had to use the mouse or arrow keys to move to the next screen. This was very cumbersome and we realised that ODK works best only on PDAs or 'smart phones'.

## 5. Concluding remarks

Our paper attempts to fill the gap in literature on research methodology with regards to use of digitised tools for data collection in large-scale surveys. Large scale surveys are resource-intensive when administered using conventional paper-pencil mode of data gathering. On the other hand, our findings show that digitised tools, particularly the free and open source ODK Toolkit, bring down the cost of data gathering, improves efficacy and can be easily employed by researchers undertaking large scale surveys even in remote areas in India.

## References

Couper, M. P. (2005). Technology trends in survey data collection. *Social Science Computer Review*, 23(4), 486-501.

Jeffrey-Coker, F., Basinger, M., & Modi, V. (2010). "Open Data Kit: Implications for the Use of Smartphone Software Technology for Questionnaire Studies in International Development." *Columbia University Mechanical Engineering Department*. Retrieved from <http://qsel.columbia.edu/assets/uploads/blog/2013/06/Open-Data-Kit-Review-Article.pdf>

Brunette, W., Sundt, M., Dell, N., Chaudhri, R., Breit, N., & Borriello, G. (2013). "Open Data Kit 2.0: Expanding and Refining Information Services for Developing Regions." In *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*, 10. ACM. Retrieved from <http://dl.acm.org/citation.cfm?id=2444790>.

Ng, W. (2015). *New Digital Technology in Education: Conceptualizing Professional Learning for Teachers*. London: Springer. (Chapter 5).

Zacharia, Z.C., Lazaridou, C., & Avraamidou, L. (2016). The use of mobile devices as means of data collection in supporting elementary school students' conceptual understanding about plants. *International Journal of Science Education*, 38(4). 596-620. DOI: 10.1080/09500693.2016.1153811

## Internet resources

Help and Download pages accessed at <https://opendatakit.org/> between November 2015 to July 2016.

## Author Biographies

**Padmini Sampath** is the Product Manager for CLIX Platform & Interactives and supports the Research team at CLIX. She may be reached at [padmini.sampath@tiss.edu](mailto:padmini.sampath@tiss.edu).

**Ashwin Nagappa** is the Lead Technologist & Co-Lead of the Technology team at CLIX. He may be reached at [ashwin.n@tiss.edu](mailto:ashwin.n@tiss.edu).

**Arundhati Roy** is currently working as a Research Associate with Research team at CLIX. She is interested in Developmental Economics and Economics of Education. She may be reached at [arundhati.roy@tiss.edu](mailto:arundhati.roy@tiss.edu).

**Ananya Chatterji** is a member of the Research team at CLIX. Her research interests include understanding the uses of language education, and how teacher education translates into practice in the classroom. She may be reached at [ananya.chatterji@tiss.edu](mailto:ananya.chatterji@tiss.edu).

**Anusha Gajinkar** has a Masters degree in Statistics and has experience as a content analyst and authoring in statistics on eLearning Platforms. She is currently working with the Research team at CLIX and may be reached at [anusha.gajinkar@tiss.edu](mailto:anusha.gajinkar@tiss.edu).

**Alpesh Gajbe** is the System Administrator at Tata Institute of Social Sciences and supports the Technology team at CLIX. He may be reached at [alpesh@tiss.edu](mailto:alpesh@tiss.edu).

**Archana Mehendale** is a Professor at Tata Institute of Social Sciences and leads the research team at CLIX. She may be reached at [archana.mehendale@tiss.edu](mailto:archana.mehendale@tiss.edu).

<sup>ii1</sup> Connected Learning Initiative (CLIX) is an initiative aimed to improve quality of teaching-learning in government high schools using technological affordances. Founded by Tata Institute of Social Sciences and Massachusetts Institute of Technology and seeded by Tata Trusts, the intervention is being rolled out in Chhattisgarh, Mizoram, Rajasthan, Telangana using

---

interactive modules developed in mathematics, science and communicative English. The baseline study was conducted as part of the larger impact evaluation study and tool development and data collection took place from October 2015- January 2016.