NIGERIA DATA COLLECTION

LESSONS ON AUTHORING MULTIPLE CHOICE SURVEY QUESTIONS

SUSTAINABLE ENGINEERING LAB
2014

SUSTAINABLE
ENGINEERING
LAB.

The following provides a few examples of the challenges encountered during the Nigeria Scale-Up project and the insights learned.

"Other" in choice lists

In multiple-choice questions, wording matters not just for the question but also for the list of answer choices. In particular, survey experts recommend choice lists where one and only one response is appropriate in any given situation. One way to make sure that *at least* one choice is always applicable, that is, to make the choice list exhaustive, is to include an *Other* choice.

```
1. What is your favorite color?
     a) Red          b) Blue          c) Green   d) Other

2. If d) Please Specify Other
```

However, there is a drawback to this method. Answers to the question *Please Specify Other* are free-text and difficult to categorize and analyze. For data analysis on large datasets, it is preferable to limit the number of free-text responses. An extreme approach to this would be to exclude the *Other* option altogether.

A more moderate and common solution to this problem is a hybrid approach. Many surveying teams include an *Other* option only during pilot stages. When they move from pilot to scale, they analyze responses from the pilot stage, make their list of choices more exhaustive, and remove the *Other* option from the final survey. This makes resulting data much easier to analyze, but it forces respondents to choose from the list of choices even if their response is not listed.

In this article, however, we highlight experience from a facilities survey in Nigeria where we have found that leaving an *Other*-like[1] option even at scale can highlight issues that go beyond exhaustiveness of choices.

Uses of an Alternate Option

In our surveying work, we have found that an alternate option like *Other* provides additional value beyond ensuring exhaustiveness. Some issues they may highlight include:

1. Misunderstanding of questions or responses by enumerators.

   Consider the following question:

   ```
   How close do you estimate that the power lines of the electricity grid
   (like NEPA or PHCN) are to this facility?

   a) Less than 500 meters
   b) Beyond 500 meters
   c) Other
   ```

[1] Note that we use *Information Not Available* / *Don't Know* instead of *Other* in our surveys. In this article, we will continue to use *Other* for brevity and clarity.
[2] The methodology used for generating the word cloud is described in the article "Monitoring the reasons for *Don't Know*", available at http://sel-columbia.github.io/formhub.R/demo/DontKnowResponses.html

By survey authoring standards, this is a good question. It asks about one thing, there isn't much room for ambiguity, and the list of options is both exhaustive and mutually exclusive. No matter where you go, the grid is either less than 500 meters away, or beyond 500 meters. Determination of whether the grid is less than or more than 500 meters away is also logistically simple, it can be perceived by the naked eye in most cases.

In practice, it turned out that a lot of enumerators selected *Other* for this question. When looking at the answer to the follow up question (*Please Specify Other*), we discovered that most answers said something like "this is a remote location" or that "community is not connected to the grid". The intended response in these cases was "Beyond 500 meters." Clearly, there was a gap in understanding between the enumerator and the survey authors.

2. Logistical / Field-Level Issues

In our surveys, enumerators can use indicate *Other* even for numerical questions (using "999"). If they choose this option, we ask an immediate follow up question asking why the answer was unknown (this is analogous to the "Please Specify Other" question). This has helped us highlight a particular issue about finding out the total number of students in a school survey. A word cloud of responses to the follow up question reveals a pattern:

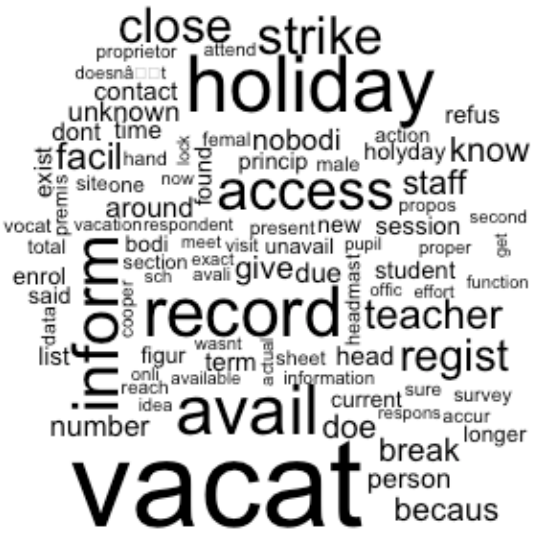
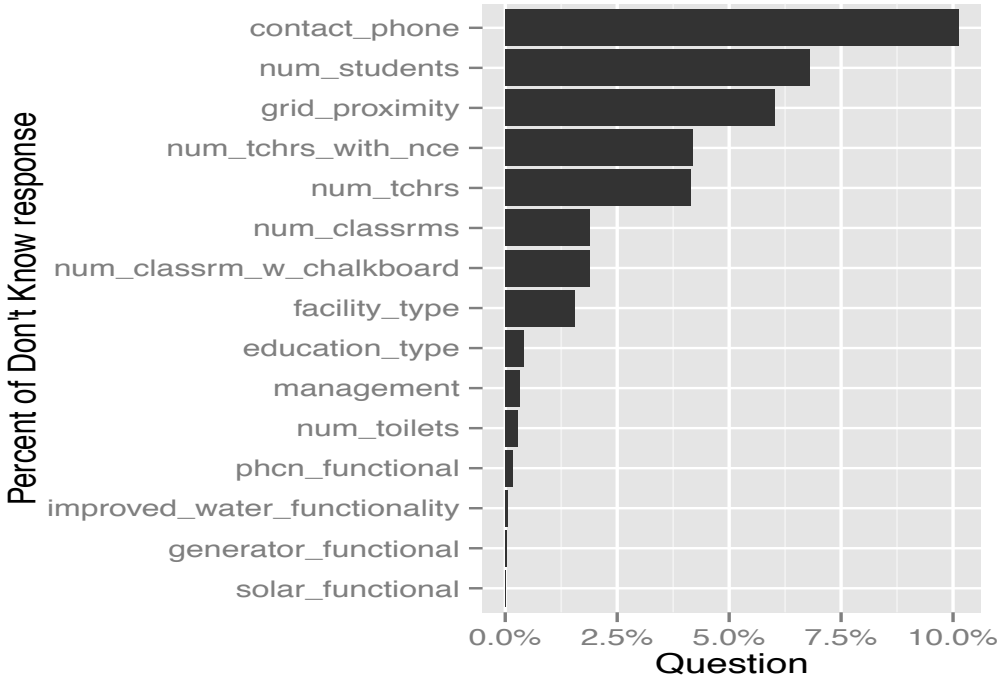


Figure 1. Word cloud for responses to why number of students was unknown. Words are stemmed; vacat corresponds to vacation. English stop words are removed, and only stems with at least 10 occurrences are shown.

It turns out that quite often, schools were on vacation or holiday, the school was closed, the teachers were on strike, or the school records were not accessible. As a result, our respondents aren't able to tell us how many students were in a school. We only discovered the extent of this issue by analyzing responses to *Please Specify Other* responses.

3. Identifying Problematic Questions

   Finally, allowing for *Other* responses allows us to find questions that are more and less difficult to gather. For example, here is a graph showing the percentage of *Other* responses by question from one stage of our survey:



   There are only three questions in which more than 5% of the responses were *Other*, and only two with more than 2.5% *Other*. First, this tells us that most of the questions and option lists are not having major issues. Additionally, it helps us determine that there are three to five questions where data quality needs further investigation, ultimately highlighting issues ranging from misunderstanding to logistics. Because we are collecting data using Android phones, sometimes we are even able to address issues while data collection is ongoing. This is invaluable to data quality.

In summary, including an *Other* option for multiple choice questions is important and effective so long as it is paired with a follow-up free-text question. This is true even for large data collection efforts where we could easily assume that there is little value in collecting thousands of free-text responses that we will not have the bandwidth to analyze. Through some relatively simple data analysis techniques[2], this *Other* data can be sorted and summarized to provide insights to pinpoint the precise issue a question is having. In addition, the follow-up question ensures that the selection of *Other* does not become an "easy way out" for enumerators to save time and speed through a survey.

[2] The methodology used for generating the word cloud is described in the article "Monitoring the reasons for *Don't Know*", available at http://sel-columbia.github.io/formhub.R/demo/DontKnowResponses.html