AUDIO CHECKS: IDENTIFYING FRAUD IN 9 COUNTRIES

Yuliya Dudaronak
ORB International
Ydudaronak@orb-international.com
OVERVIEW

Issue
Interview falsification has been long established as a significant issue in the collection of survey data. Even a modest number of falsified cases can have a significant impact on analysis of the dataset, and some types of falsification and other poor interviewing practices can be particularly difficult to catch.

Audio Checking
Audio checking records a sample of interview questions to be listened to and checked for falsification or poor interviewing.

Results
- Detected data fabrication in 7% to 24% of our surveys.
- Undetected fabrication would have introduced significant bias in our analyses.
- The audio check performs well compared to more traditional methods of detecting fabrication.
ACHIEVED CAPI SAMPLE

All Muslim
Male/Female
All education levels

Afghanistan: 981
Tunisia: 2,484
Mali: 2,525
Guinea: 1,955
Nigeria: 2,784
Niger: 2,525
Chad: 3,129
Pakistan: 2,820
Malaysia: 2,413
QUALITY CONTROL IS COMPRISED OF MULTIPLE CHECKS

Supervisor checks
Audio checks
Geolocation checks
Process oriented checks
TRADITIONAL APPROACHES AND LIMITATIONS

**Backchecking**
Supervisors contacting, in person or by telephone, a sample (generally between five and fifteen percent) of each interviewer’s respondents and confirming that an interview took place and that the answers were recorded correctly.

Only addresses complete fraud, does not address skipped question

**Statistical Methods**
Use of interview data and metadata to identify cases or interviewers who are likely to be false based on duplicated/matching interviews, contradictory time, location, and lengths of interviews, etc.

Conducted after fieldwork completion, long time and costly to refield

**Detecting Cheating Interviewers**
Identify interviewers who are likely to be fabricating data thought analysis of skipping/extreme answers.skipped questions patterns, patterns of rare combinations in participants’ responses, non-response, straightlining, etc.

Limited in their ability to highlight specific cases as being likely to be fabricated, because of the often-unpredictable nature of genuine responses to surveys.
4 question recorded using Survey to Go software

Data downloaded daily

Trained QC staff listens to audio

Failed cases deleted and immediately refielded
AUDIO CHECK SYSTEM

Bad audio captures (fail):  
- Snippets of sound from interviewer/respondent
- Question mumbled to the point of inaudibility
- Radio or TV broadcast, religious service, etc., plays in background
- “Dead air,” ambient background noise, wind/rain
- Interactions not belonging to interviewer or respondent

Good audio captures (non-English):  

Good audio captures (English):  

RESULTS
Overall, 15% of data was deleted from our sample based on audio – 3249 cases out of 21628.
SURVEY LENGTH

Mean duration:

Accepted: 30:33 min
Failed: 23:46 min
Difference of Means
Calculated the difference in means between passed and failed surveys, controlling for country-specific variation in means.

Variance Comparison
Compared the ratio of variance of passed vs failed interviews - previous research indicated that fabricated surveys have on average lower variance.

Extreme Answers
Rate of recoded extreme responses – i.e., on a scale from 1 to 7, responding 1, 2, 6, or 7. Individuals who fabricate surveys underestimate the number of times participants provide extreme answers.

Percent ‘Don’t know’ and ‘Refused’
Compared the rate of item non-response – ‘Don’t know’ and ‘Refused’ answers.

Conflicting answers rate
Compared the rate of answers that have logical inconsistencies. Individuals who fabricate surveys underestimate the number of times participants provide conflicting answers.

Rate of ‘other’ answers
Proportion of questions where ‘other’ is an option which were answered with ‘other’, triggering an open-ended entry.
DIFFERENCE OF MEANS

Method

1. Standardized each variable across the entire sample - the estimated difference is in terms of the (overall) standard deviation for that variable

2. Regression with a set of dummy variables for each country to account for the possibility of different means across countries, plus a dummy variable for "failed"

3. Graphed the difference in the mean for failed interviews compared with non-failed (as an effect size)

Results

1. Fraudulent interviews have significantly different means in our survey [Statistically significant difference for 32 of 63 variables (51%) at 95% confidence interval]

2. Statistical differences run across different topics of the survey, from media use to attitudes towards the government
VARIANCE COMPARISON

Method

1. Heteroskedastic regression model of each DV on dummy variables per country, estimating the impact of failed-vs-accepted on the residual variance. This has the effect of calculating the variances with respect to the country-specific means.

2. Because the test is done on the ratio of the variances, the null hypothesis is one (not zero).

Results

1. Fraudulent interviews have significantly different distribution of answers [Statistically significant variance difference for 26 of 63 variables (41%) at 95% confidence interval].

2. Statistical differences run across different topics of the survey, from media use to attitudes towards the government.
TESTS FOR ANSWER PATTERNS

Fraudulent interviews display significantly different answer patterns on all traditional tests

- Interviewers underestimate the number of times people agree with all items in the battery (straightlining), give extreme or conflicting answers, and are less likely to select ‘other’ as an answer option.
- Interviewers who fake data are more likely to select ‘Don’t know’ or ‘Refused’.

Mean differences for failed interviews

- Straightline rate
- Extreme answer rate (of 48)
- DK and RF rate
- Conflicting answer rate (of 4)
- Other rate (of 4)
- Interview duration (100s of minutes)

Difference of mean between failed and good interviews. Using unstandardized variables, with 95 percent confidence intervals. p<0.05 for 6 of 6 comparisons (100.0%)
Effect size of difference in means between full sample and 'good' sample with 95 percent confidence intervals. p<0.05 for 239 of 567 comparisons (42.2%)
IMPACT OF FAILED INTERVIEWS ON ESTIMATED MEANS

Effect size of difference in means between full sample and 'good' sample with 95 percent confidence intervals. p<0.05 for 239 of 567 comparisons (42.2%)
Effect size of difference in means between full sample and 'good' sample with 95 percent confidence intervals. p<0.05 for 239 of 567 comparisons (42.2%)
INTERVIEWER CHARACTERISTICS
Limited data is available on the effect of interviewer characteristics on failure rates.

This graphs shows average failure rates for each category (e.g., male failure rate is 17.5%, female is 11.4%, etc.) compared to the overall average failure rate.
TAKE AWAY

We also delete cases based on other quality checks (i.e. GPS - about 2% overall), but **audio checks are key to high quality data.**

- Backchecks failed to identify any of these cases as fraudulent
- Observed interviews – no difference between failed and passed (i.e. observed interviews have the same likelihood of failing as not observed interviews)
BEST PRACTICES FOR AUDIO CHECKS

Carefully consider which questions to record, in particular:

- The location in the questionnaire;
- Sensitivity;
- Filters used in a survey;
- Applicability to all respondents.

Decide ahead of time on the proportion of audio checks, the standard for good vs bad, and how you going to handle feedback to local

Increase interviewer training