Appendix_Supplemental Material

The Interviewer Performance Profile (IPP): A Paradata-Driven Tool for Monitoring and Managing Interviewer Performance

Heidi M. Guyer, RTI International, Research Triangle Park, NC
Brady T. West, Survey Research Center, University of Michigan, Ann Arbor, MI
Wen Chang, Survey Research Center, University of Michigan, Ann Arbor, MI

Key Performance Indicator (KPI) Definitions

- Following are the definitions for each of the key performance indicators.
    - **Hours:** Total number of hours for data collection-related tasks, including screening, interviewing, travel, administrative tasks, and dealing with technical issues.
    - **Percentage of production hours:** The percentage of total hours spent on screening and interviewing.
    - **Hours per interview (HPI):** Total number of hours reported by the interviewer divided by the total number of main interviews completed.
    - **Screening interviews:** Total number of completed screeners, excluding screeners completed by proxy.
    - **Main interviews:** Total number of completed main interviews.
    - **Screening completion rate:** The percentage of sampled households that were finalized.
    - **Main completion rate:** The percentage of eligible households that were finalized.
    - **Screening response rate:** Total number of completed screeners divided by the total number of sampled households, excluding non-sample cases

- **Main interview response rate:** Total number of completed main interviews divided by total number of eligible households, excluding non-sample cases
- **Eligibility rate:** number of eligible households divided by the number of completed screeners.

PAIP Scores

- **Interview PAIP:** To take into account the difficulty of the assigned sample, selected paradata, in addition to other auxiliary variables available for both respondents and non-respondents, are used in a discrete-time hazard model to estimate the interview response propensity at the immediate subsequent contact. The contact-level PAIP score is calculated by subtracting the response propensity from the actual outcome. For example, if the predicted response propensity at next contact for an active sample case is 0.2, then the contact-level PAIP score for that case would be 1-0.2=0.8 for a successful interview or 0-0.2=-0.2 for an unsuccessful interview. This approach therefore gives large credit when obtaining success on very difficult cases, and only a small penalty given failure with such cases. For each interviewer, the contact-level PAIP scores for all active cases are averaged over all contacts to compute the Interview PAIP score.
- **Eligibility PAIP:** Information from the sampling frame, such as census division, sampling domain based on information regarding the race and ethnicity distribution , the estimated eligibility rate based on American Community Survey (ACS) data, and commercial data matched to the

addresses are used to estimate the eligibility propensity (i.e., the probability of someone between the ages of 15 and 49 being present) for each selected address. The address-level PAIP score is calculated as the difference between the outcome of the screening interview (1 for eligible and 0 for ineligible) and the estimated eligibility propensity. For each interviewer, the Eligibility PAIP is the average of the address-level scores over all cases screened by each interviewer.

- **Contact PAIP:** The probability of achieving contact at each contact attempt is predicted by Census region, urbanicity of the area, several interviewer observations made during address listing, and the time window of the attempt. The attempt-level difference between the outcome of the attempt (1 for contact and 0 for no contact) and the predicted contact propensity is calculated for each attempt. These attempt-level scores are averaged over all attempts made by each interviewer as the Contact PAIP score.

Data Quality Indicators

- These three components are described in more detail below.
  - **Too fast:** high component loadings of short field time and frequently backing up within the survey (potentially to change or review a response), and moderate loadings of frequently closing or skipping over error checks.
  - **High error checks:** high component loadings of frequent error suppression and moderate loadings of frequent closing and skipping over error checks.
  - **High don't know or refused responses:** high component loadings of frequent don't know and refusal responses on survey items.

Scores on each component are computed for each interviewer based on the component loadings, and these scores were then standardized to create z-scores. Those interviewers with z-scores greater than two standard deviations were considered having poor data quality (e.g., unusually high error checks) whereas those with z-scores two standard deviations below were considered having good data quality (e.g., fewer error checks). These three data quality indicators are calculated every two weeks based on the interviews completed in the prior two weeks to help monitor changes over time.

Data Set Balance Indicators

- Indicators that were evaluated at the interviewer level included:
    - The percentage of households with children aged 14 or younger, based on data from the screening interview.
    - The percentage of selected respondents who are likely sexually active based on an interviewer observation (West & Kreuter, 2015).
    - Main interview response rates for demographic subgroups traditionally at risk of low response rates (e.g., Hispanics).

Paradata Tables

Sample table: The tables included a table with sample line level details (ID, project name, assigned interviewer, current disposition, date last attempted), all contact attempt level data for each sample line (date of call, call window, result code, flags set for resistance and contact, appointment date and time), the transfer history for each case, employee level data (interviewer ID, start date, geographic area of work, experience level, supervisor), and

employee timesheet data (hours by work type and by date). Age, sex and race/ethnicity of the selected respondents were extracted from screener interview data.

Effort table: As an indicator of effort, hours were categorized by work type (administrative, travel, screening, production) and aggregated by interviewer and by day. Hours per day were summed to calculate cumulative hours by interviewer as well.

Productivity table: Productivity measures were summarized for each interviewer at the daily and cumulative levels and included:

- **Result type**: counts of each result type (interview, refusal, no contact, no attempt, and non-sample) for screening and main stages of the data collection.
- **Calls by outcome type**: the total number of calls and the count of calls by outcome type (interview, appointment, missed appointment, non-interview, whether contacted, whether resisted) was calculated overall and by subgroup for each interviewer, for the current day as well as cumulatively.
- **Status of outstanding cases**: the status of outstanding, or non-final, sample was evaluated using the call information through the end of each day. Given that sample can be transferred from one interviewer to another, it is important to have an accurate reflection of the current sample for each interviewer as well as their own work to date.

Using the IPP

Key Performance Indicators (KPIs)

The KPIs for sixteen interviewers are displayed in Table 1, with each row representing one interviewer. The KPIs displayed include the total number of hours worked, the percentage of administrative hours, HPI, the total number of screener and main interviews completed, and completion, response and eligibility rates. We observe

that Interviewer 1 has completed an average number of screener and main interviews, 26 and 7 respectively, based on the yellow shading. However, Interviewer 1's HPI, an indicator of efficiency, is green, indicating a good HPI compared to other interviewers. We also see that Interviewer 1 is struggling in screener completion and response rates, which indicates an area for additional coaching and improvement needed. In comparison, Interviewer 2 has only completed 13 screener interviews and 0 main interviews, and each KPI is shaded in red, which indicates that Interviewer 2 is one of the poorest performing interviewers in this area of productivity. Additionally, Interviewer 2 is performing poorly in screener and main completion rates and response rates. Across all sixteen interviewers, we observe that this group is doing well in total hours and the percent of administrative hours, but not doing as well in screening completion and response rates. These indicate areas to acknowledge as positive performance as well as areas where additional coaching and monitoring may be needed.

Propensity-Adjusted Interviewer Performance (PAIP) Scores

The PAIP scores for sixteen interviewers are shown in Table 2. Each row represents one interviewer. First, discrepancies between respondents and non-respondents in the percentage with young kids and the percentage of adults perceived to be sexually active are shown. Smaller discrepancies are considered positive performance whereas increased discrepancies indicate poor performance. Interviewer 8 is performing well on both indicators. These metrics are followed by five PAIP scores: screener interview completion, main interview completion, eligibility rates, screener contact rates and main contact rates. Here we see that more interviewers are performing well on screener interview completion rates than main interview

completion rates, indicating an area for potential intervention. Interviewer 3 is performing well with contact rates, but not with interview completion rates. Additional conversations with Interviewer 3 may be needed to better understand the barriers to completing screening and main interviews among those she is able to contact. Additional training may also be required to aide Interviewer 3 in gaining skills to "seal the deal".

Data Quality Indicators

Table 3 displays the z-scores for the three primary data quality indicators identified: going too fast, frequent error checks, and high frequency of don't know or refused responses. At a quick glance, we see that these sixteen interviewers are generating fewer error checks than others on average. However, three interviewers, numbers 5, 6 and 9, have higher frequencies of don't know and refusal responses. Interviewers 5 and 9 seem to be struggling in two areas: going too fast and entering many don't know and refused responses. These two interviewers will require retraining and may be temporarily stopped from data collection in order to address the data quality concerns. Ongoing performance issues related to poor data quality may indicate that they are not a good fit for the study. These indicators are updated every two weeks, and thus observed changes in performance may require several weeks while the interviewer is retrained and then conducts additional interviews.

Data Set Balance Indicators

Four indicators of data set balance are shown in Table 4: teen response rates, adult male Hispanic (AMH) response rates, and the percentage of active (non-final) screener and main interview sample that indicated reluctance to participate on a previous contact attempt. Among these sixteen interviewers, we observe that most

have low reluctance rates based on their current sample. While this appears positive and could mean that they are achieving cooperation at a higher rate than their peers, it could be also be attributed to other factors that should be considered. For example, they may have a low number of active lines compared to other interviewers, or they may have transferred all of their reluctant cases to other more experienced interviewers. If neither of these are the case, it may be helpful to have this group of interviewers provide tips and strategies to interviewers that are struggling in this area of performance. Interviewer 2 appears to be having difficulty completing interviews with teenagers and with AMH respondents. Additional guidance could be provided to Interviewer 2 to help improve their performance with respondents having these characteristics.

Supplemental Table 1. IPP: Key Performance Indicators (KPIs)

						More in the good direction								More in the bad direction		
Interviewer	Data Cumulated To	Note	Hour Commitment	# PSU Attempted	% of Called to Main PSU	Hours (SMTAC)	% of scrn and main hours	% of admin hours	HPI	Screener IW	Main IW	Screener Completion Rate	Main Completion Rate	Scrn RR (unweighted)	Main RR (unweighted)	Eligibility Rates
1	07/07/2019	Y7	30	1	100%	63	57%	15%	9.0	26	7	21%	35%	18%	35%	77%
2	07/07/2019	Y7	30	1	100%	46	45%	16%	.	13	0	21%	0%	19%	0%	62%
3	07/07/2019	Y7	30	1	100%	52	62%	14%	17.4	28	3	37%	18%	36%	18%	55%
4	07/07/2019	Y7	30	1	100%	115	70%	2%	6.4	54	18	29%	69%	27%	69%	48%
5	07/07/2019	Y7	30	1	100%	79	68%	4%	9.9	39	8	37%	62%	35%	62%	26%
6	07/07/2019	Y7	30	1	100%	89	45%	6%	.	4	.	17%	.	17%	.	0%
7	07/07/2019	Y7	30	1	100%	75	58%	7%	25.1	25	3	37%	18%	36%	18%	31%
8	07/07/2019	Y7	30	1	100%	68	65%	13%	13.5	34	5	33%	31%	33%	31%	56%
9	07/07/2019	Y7	30	1	100%	85	49%	4%	10.7	46	8	26%	40%	22%	40%	43%
10	07/07/2019	Y7	30	1	100%	110	63%	10%	13.7	42	8	60%	42%	59%	42%	48%
11	07/07/2019	Y7	30	1	100%	116	71%	5%	10.6	90	11	46%	26%	38%	26%	46%
12	07/07/2019	Y7	30	1	100%	155	79%	4%	7.1	74	22	48%	100%	42%	100%	46%
13	07/07/2019	Y7	30	1	100%	121	71%	10%	8.1	88	15	43%	34%	41%	34%	51%
14	07/07/2019	Y7	30	1	100%	90	50%	9%	11.3	37	8	39%	36%	34%	36%	59%
15	07/07/2019	Y7	30	1	100%	71	65%	12%	14.3	45	5	34%	22%	29%	14%	53%
16	07/07/2019	Y7	30	1	100%	121	64%	10%	10.1	58	12	44%	36%	39%	33%	54%

Supplemental Table 2. IPP: PAIP Indicators

Interviewer	% Kids Discrepancy (R-NR)	% Sexual Active Discrepancy (R-NR)	PAIP - Screener Completion	PAIP Main Completion	PAIP - Eligibility Rates	PAIP - Screener Contact Rates	PAIP - Main Contact Rates
1	2%	-7%	-17%	10%	12%	3%	46%
2	.	.	13%	-15%	-8%	-7%	44%
3	19%	-45%	-18%	-10%	0%	14%	37%
4	15%	8%	15%	14%	0%	-4%	40%
5	-30%	-13%	15%	33%	-24%	-4%	13%
6	.	.	-23%	.	-29%	-10%	.
7	-2%	0%	0%	-5%	-30%	2%	40%
8	-24%	-53%	-10%	-19%	-4%	18%	11%
9	-4%	0%	-4%	15%	-4%	4%	37%
10	-33%	-13%	-13%	4%	-1%	27%	39%
11	-6%	-13%	-2%	-9%	2%	-3%	10%
12	.	.	6%	39%	-10%	3%	6%
13	17%	7%	15%	0%	2%	11%	45%
14	7%	16%	10%	13%	-5%	-5%	33%
15	-5%	-24%	-9%	-11%	0%	17%	46%
16	8%	0%	-10%	-4%	-4%	-6%	13%

Note: In the table above, completion indicates interview completion rather than case completion.

Supplemental Table 3: IPP: Data Quality Indicators

Interviewer	Data Quality - Too Fast	Data Quality - Many Error Checks	Data Quality - Many DK/RF
1	-0.3	1.2	0.6
2	3.0	0.1	0.1
3	-0.2	-1.2	-0.7
4	-0.3	-0.8	-0.3
5	1.2	-0.9	3.4
6	0.1	-0.8	3.2
7	0.3	2.0	-0.4
8	0.1	1.5	-0.8
9	1.6	-0.7	3.5
10	-0.4	0.8	-0.1
11	-0.6	-1.3	0.0
12	1.5	-0.7	-0.4
13	0.5	0.5	-0.7
14	-0.3	-0.4	-0.2
15	-0.6	0.8	0.5
16	0.2	1.6	-0.5

Supplemental Table 4: IPP: Data Set Balance Indicators

Interviewer	# of Teen Lines Assigned	Teen Main RR	# of Male Hispanic Adults	Male Hispanic Adults Main RR	# of Active Ever Contacted Screener	% Ever Resisted Active Screener	# of Active Main	% Ever resisted Active Main
1	3	33%	.	.	27	37%	13	0%
2	2	0%	4	0%	1	0%	8	0%
3	7	43%	.	.	13	8%	14	0%
4	5	80%	3	100%	3	33%	8	0%
5	3	100%	3	33%	0	.	5	0%
6	.	.	.	.	12	50%	.	.
7	2	0%	.	.	17	24%	14	7%
8	2	0%	2	50%	16	44%	11	0%
9	4	50%	.	.	11	0%	12	0%
10	6	67%	.	.	23	61%	11	9%
11	10	30%	8	13%	15	40%	31	13%
12	2	100%	1	100%	0	.	0	.
13	12	25%	7	71%	5	80%	29	7%
14	6	33%	.	.	4	25%	14	7%
15	7	43%	4	0%	17	41%	28	18%
16	10	30%	.	.	23	26%	23	9%