INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER
SURVEY RESEARCH OPERATIONS
UNIVERSITY OF MICHIGAN

Using Paradata to Develop and Implement an Interviewer Performance Profile for Monitoring and Evaluating Interviewer Performance

Wen Chang, Heidi Guyer, Brady T. West

Institute for Social Research, University of Michigan

Background

New field management tools developed for a national survey

Data monitoring

- **Project level:** Selection and display of key indicators in a dashboard view
- **Interviewer level:** Performance profile

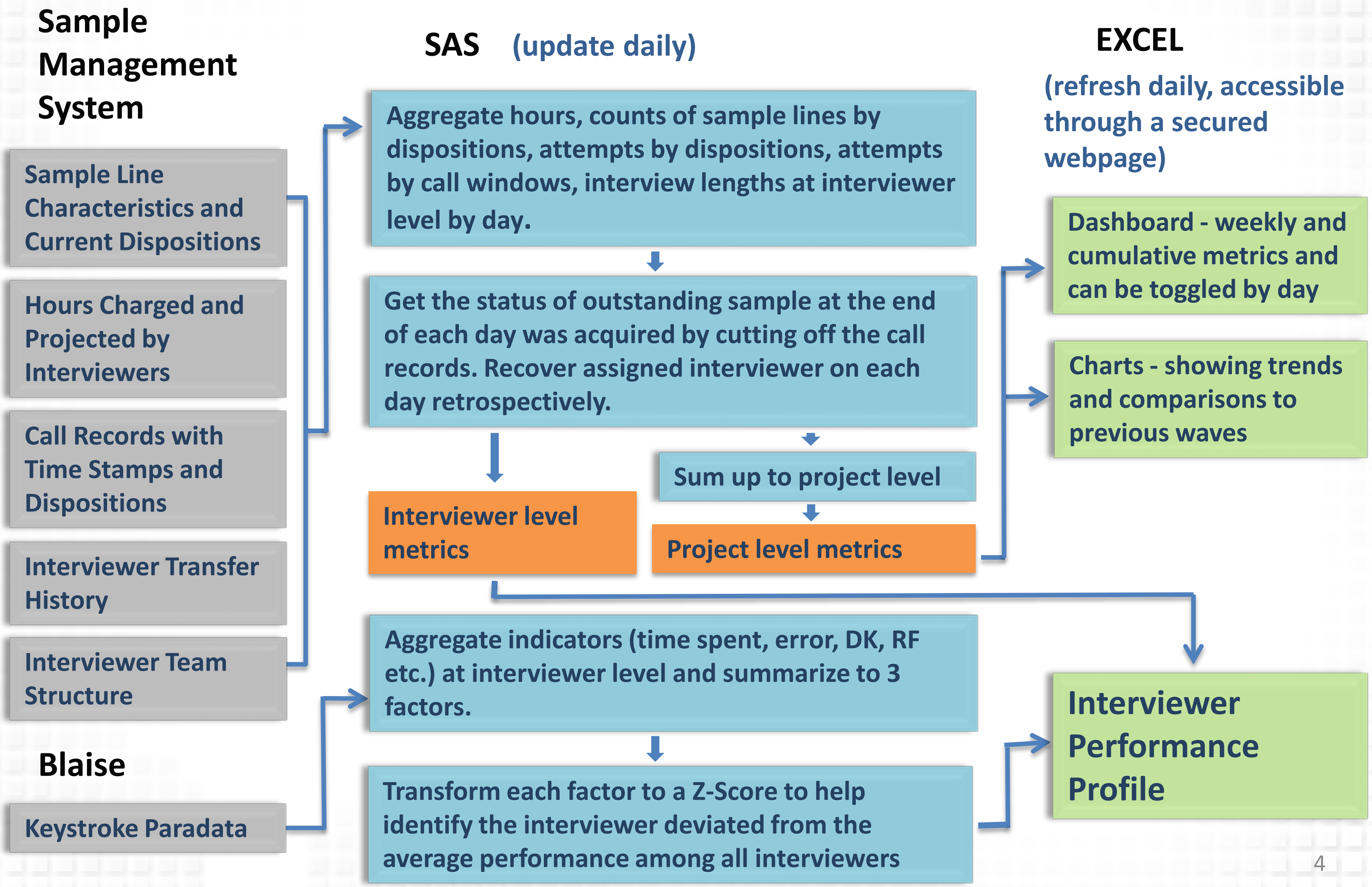
Paradata sources

- **Production monitoring:** Sample management system
- **Quality control:** Interview keystroke paradata

Interviewer Performance Profile

- Summarizes information at the interviewer level
- Uses heat maps to identify positive versus negative performance at a quick glance
- Integrates most up to date data on effort, productivity and quality
- Can be assessed based on pre-determined time frame (full year, cumulative for the month, etc.)
- An easy tool for methodologists and managers to identify individual areas of concern plus tracking progress after interventions (re-training)

Data Flow



Sample Management System

Sample Line Characteristics and Current Dispositions

Hours Charged and Projected by Interviewers

Call Records with Time Stamps and Dispositions

Interviewer Transfer History

Interviewer Team Structure

Blaise

Keystroke Paradata

SAS (update daily)

Aggregate hours, counts of sample lines by dispositions, attempts by dispositions, attempts by call windows, interview lengths at interviewer level by day.

Get the status of outstanding sample at the end of each day was acquired by cutting off the call records. Recover assigned interviewer on each day retrospectively.

Interviewer level metrics

Sum up to project level

Project level metrics

Aggregate indicators (time spent, error, DK, RF etc.) at interviewer level and summarize to 3 factors.

Transform each factor to a Z-Score to help identify the interviewer deviated from the average performance among all interviewers

EXCEL

(refresh daily, accessible through a secured webpage)

Dashboard - weekly and cumulative metrics and can be toggled by day

Charts - showing trends and comparisons to previous waves

Interviewer Performance Profile

INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER
SURVEY RESEARCH OPERATIONS
UNIVERSITY OF MICHIGAN

Interviewer Performance Profile – Overall View

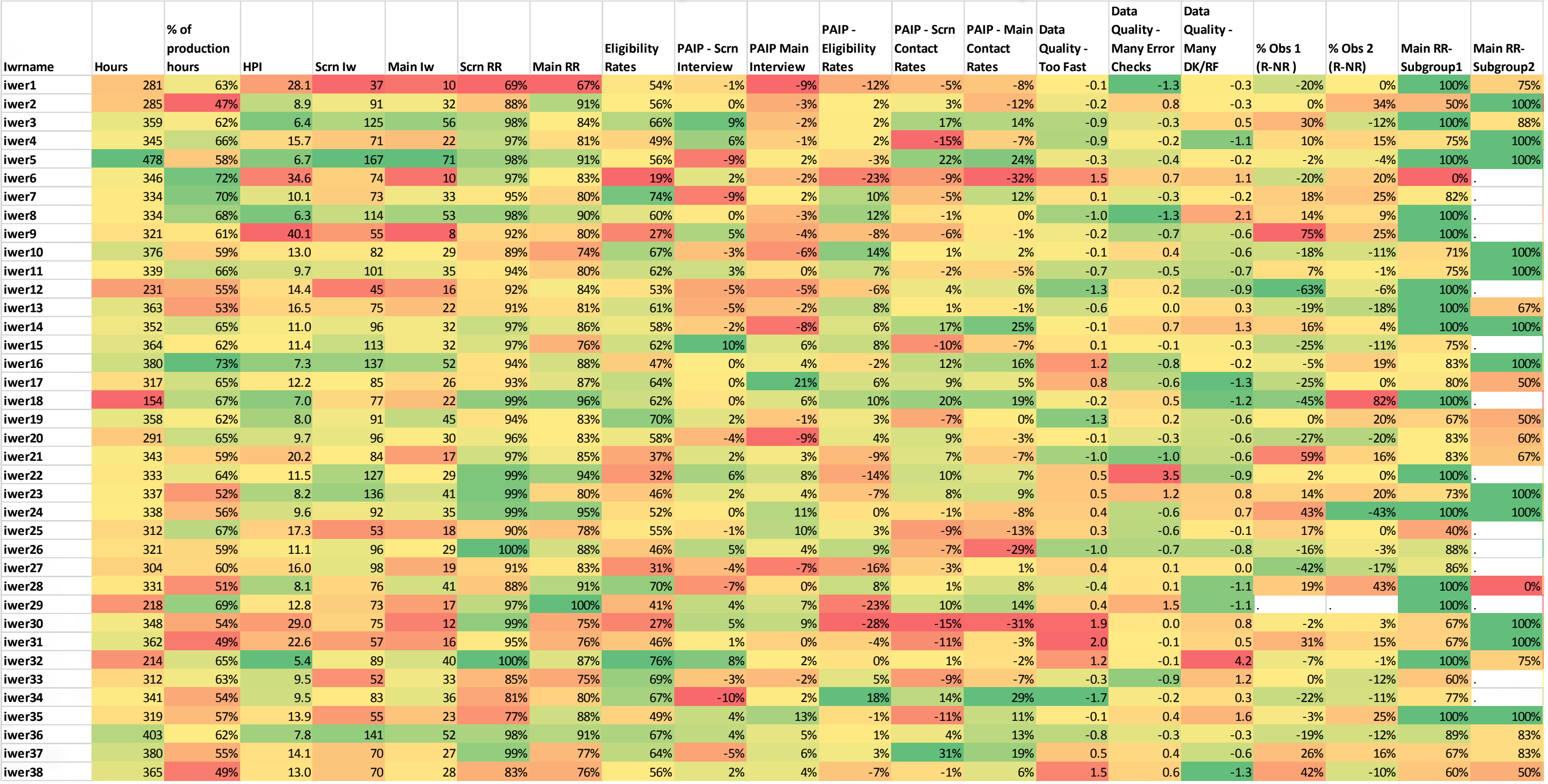
Each row shows the metrics for one interviewer

Key Performance Indicators

PAIP Indicators

Data Quality Indicators

Data Set Balance Metrics



Iwrname	Hours	% of production hours	HPI	Scrn Iw	Main Iw	Scrn RR	Main RR	Eligibility Rates	PAIP - Scrn Interview	PAIP Main Interview	PAIP - Eligibility Rates	PAIP - Scrn Contact Rates	PAIP - Main Contact Rates	Data Quality - Too Fast	Data Quality - Many Error Checks	Data Quality - Many DK/RF	% Obs 1 (R-NR)	% Obs 2 (R-NR)	Main RR-Subgroup1	Main RR-Subgroup2
iwer1	281	63%	28.1	37	10	69%	67%	54%	-1%	-9%	-12%	-5%	-8%	-0.1	-1.3	-0.3	-20%	0%	100%	75%
iwer2	285	47%	8.9	91	32	88%	91%	56%	0%	-3%	2%	3%	-12%	-0.2	0.8	-0.3	0%	34%	50%	100%
iwer3	359	62%	6.4	125	56	98%	84%	66%	9%	-2%	2%	17%	14%	-0.9	-0.3	0.5	30%	-12%	100%	88%
iwer4	345	66%	15.7	71	22	97%	81%	49%	6%	-1%	2%	-15%	-7%	-0.9	-0.2	-1.1	10%	15%	75%	100%
iwer5	478	58%	6.7	167	71	98%	91%	56%	-9%	2%	-3%	22%	24%	-0.3	-0.4	-0.2	-2%	-4%	100%	100%
iwer6	346	72%	34.6	74	10	97%	83%	19%	2%	-2%	-23%	-9%	-32%	1.5	0.7	1.1	-20%	20%	0%	.
iwer7	334	70%	10.1	73	33	95%	80%	74%	-9%	2%	10%	-5%	12%	0.1	-0.3	-0.2	18%	25%	82%	.
iwer8	334	68%	6.3	114	53	98%	90%	60%	0%	-3%	12%	-1%	0%	-1.0	-1.3	2.1	14%	9%	100%	.
iwer9	321	61%	40.1	55	8	92%	80%	27%	5%	-4%	-8%	-6%	-1%	-0.2	-0.7	-0.6	75%	25%	100%	.
iwer10	376	59%	13.0	82	29	89%	74%	67%	-3%	-6%	14%	1%	2%	-0.1	0.4	-0.6	-18%	-11%	71%	100%
iwer11	339	66%	9.7	101	35	94%	80%	62%	3%	0%	7%	-2%	-5%	-0.7	-0.5	-0.7	7%	-1%	75%	100%
iwer12	231	55%	14.4	45	16	92%	84%	53%	-5%	-5%	-6%	4%	6%	-1.3	0.2	-0.9	-63%	-6%	100%	.
iwer13	363	53%	16.5	75	22	91%	81%	61%	-5%	-2%	8%	1%	-1%	-0.6	0.0	0.3	-19%	-18%	100%	67%
iwer14	352	65%	11.0	96	32	97%	86%	58%	-2%	-8%	6%	17%	25%	-0.1	0.7	1.3	16%	4%	100%	100%
iwer15	364	62%	11.4	113	32	97%	76%	62%	10%	6%	8%	-10%	-7%	0.1	-0.1	-0.3	-25%	-11%	75%	.
iwer16	380	73%	7.3	137	52	94%	88%	47%	0%	4%	-2%	12%	16%	1.2	-0.8	-0.2	-5%	19%	83%	100%
iwer17	317	65%	12.2	85	26	93%	87%	64%	0%	21%	6%	9%	5%	0.8	-0.6	-1.3	-25%	0%	80%	50%
iwer18	154	67%	7.0	77	22	99%	96%	62%	0%	6%	10%	20%	19%	-0.2	0.5	-1.2	-45%	82%	100%	.
iwer19	358	62%	8.0	91	45	94%	83%	70%	2%	-1%	3%	-7%	0%	-1.3	0.2	-0.6	0%	20%	67%	50%
iwer20	291	65%	9.7	96	30	96%	83%	58%	-4%	-9%	4%	9%	-3%	-0.1	-0.3	-0.6	-27%	-20%	83%	60%
iwer21	343	59%	20.2	84	17	97%	85%	37%	2%	3%	-9%	7%	-7%	-1.0	-1.0	-0.6	59%	16%	83%	67%
iwer22	333	64%	11.5	127	29	99%	94%	32%	6%	8%	-14%	10%	7%	0.5	3.5	-0.9	2%	0%	100%	.
iwer23	337	52%	8.2	136	41	99%	80%	46%	2%	4%	-7%	8%	9%	0.5	1.2	0.8	14%	20%	73%	100%
iwer24	338	56%	9.6	92	35	99%	95%	52%	0%	11%	0%	-1%	-8%	0.4	-0.6	0.7	43%	-43%	100%	100%
iwer25	312	67%	17.3	53	18	90%	78%	55%	-1%	10%	3%	-9%	-13%	0.3	-0.6	-0.1	17%	0%	40%	.
iwer26	321	59%	11.1	96	29	100%	88%	46%	5%	4%	9%	-7%	-29%	-1.0	-0.7	-0.8	-16%	-3%	88%	.
iwer27	304	60%	16.0	98	19	91%	83%	31%	-4%	-7%	-16%	-3%	1%	0.4	0.1	0.0	-42%	-17%	86%	.
iwer28	331	51%	8.1	76	41	88%	91%	70%	-7%	0%	8%	1%	8%	-0.4	0.1	-1.1	19%	43%	100%	0%
iwer29	218	69%	12.8	73	17	97%	100%	41%	4%	7%	-23%	10%	14%	0.4	1.5	-1.1	.	.	100%	.
iwer30	348	54%	29.0	75	12	99%	75%	27%	5%	9%	-28%	-15%	-31%	1.9	0.0	0.8	-2%	3%	67%	100%
iwer31	362	49%	22.6	57	16	95%	76%	46%	1%	0%	-4%	-11%	-3%	2.0	-0.1	0.5	31%	15%	67%	100%
iwer32	214	65%	5.4	89	40	100%	87%	76%	8%	2%	0%	1%	-2%	1.2	-0.1	4.2	-7%	-1%	100%	75%
iwer33	312	63%	9.5	52	33	85%	75%	69%	-3%	-2%	5%	-9%	-7%	-0.3	-0.9	1.2	0%	-12%	60%	.
iwer34	341	54%	9.5	83	36	81%	80%	67%	-10%	2%	18%	14%	29%	-1.7	-0.2	0.3	-22%	-11%	77%	.
iwer35	319	57%	13.9	55	23	77%	88%	49%	4%	13%	-1%	-11%	11%	-0.1	0.4	1.6	-3%	25%	100%	100%
iwer36	403	62%	7.8	141	52	98%	91%	67%	4%	5%	1%	4%	13%	-0.8	-0.3	-0.3	-19%	-12%	89%	83%
iwer37	380	55%	14.1	70	27	99%	77%	64%	-5%	6%	3%	31%	19%	0.5	0.4	-0.6	26%	16%	67%	83%
iwer38	365	49%	13.0	70	28	83%	76%	56%	2%	4%	-7%	-1%	6%	1.5	0.6	-1.3	42%	-10%	60%	50%

Color coding: Green means good performance and red means poor performance

The cell that holds median among all interviewers is highlighted in yellow. If a larger value of an indicator means better performance, e.g. response rate, then the cell that holds maximum value is highlighted in green and the cells that holds minimum is highlighted in red. All other cells are colored proportionally. Green and red are used the other way around if a smaller value of an indicator means better performance.

Monitoring Interviewer Performance: Key Performance Indicators (KPIs)

KPIs:

- Effort:
  - Total hours
  - % of production hours (hours for screening and completing main interview)
- Hours Per Interview (HPI)
- Interview Yield
- Response Rates
- Eligibility Rates

Monitoring Interviewer Performance: PAIP Indicator

PAIP: **P**ropensity-**A**djusted **I**nterviewer **P**erformance

- PAIP scores created as performance indicators
- Account for difficulty and/or sample characteristics
- Evaluates the effectiveness of the interviewer by incorporating *difficulty* of the interviewing task at the contact level
- Eligibility propensity and contact rate propensity are estimated by separate models. PAIP scores are calculated in the same fashion at line and attempt level for eligibility rate and contact rate, respectively

Monitoring Interviewer Performance: PAIP Indicator

- 3 steps:

  1. Available paradata are used to estimate the propensity that the next contact with the active case will generate an interview

  2. Calculate response propensity:
     – A successful interview on the next contact => 1 - estimated response propensity
     – A non-successful interview on the next contact => 0 - estimated response propensity

  3. For each interviewer, the contact-level scores are averaged over all contacts

- Gives large credit when obtaining success on very difficult cases, and only a small penalty given failure with such cases. The other way around for easy cases.

  For example, if estimated response propensity = 0.8:
  – A successful interview = 1 - 0.8 = 0.2 PAIP score
  – An unsuccessful interview = 0 - 0.8 = -0.8 PAIP score

Monitoring Interviewer Performance: PAIP Indicator

PAIP scores used to evaluate performance on:

- Completing interviews
- Identifying eligible households
- Achieving contacts

Monitoring Interviewer Performance: PAIP Indicator

Iwrname	Scrn RR	Main RR	Eligibility Rates	PAIP - Scrn Interview	PAIP Main Interview	PAIP - Eligibility Rates	PAIP - Scrn Contact Rates	PAIP - Main Contact Rates
iwer1	**69%**	67%	54%	**-1%**	-9%	-12%	-5%	-8%
iwer2	88%	91%	56%	0%	-3%	2%	3%	-12%
iwer24	99%	95%	52%	0%	11%	0%	-1%	-8%
iwer25	90%	78%	55%	-1%	10%	3%	-9%	-13%
iwer26	100%	88%	**46%**	5%	4%	**9%**	-7%	-29%
iwer27	91%	83%	31%	-4%	-7%	-16%	-3%	1%
iwer28	88%	91%	70%	-7%	0%	8%	1%	8%
iwer29	97%	100%	41%	4%	7%	-23%	10%	14%
iwer30	99%	**75%**	27%	5%	**9%**	-28%	-15%	-31%
iwer31	95%	76%	46%	1%	0%	-4%	-11%	-3%
iwer32	100%	87%	**76%**	8%	2%	**0%**	1%	-2%
iwer33	85%	75%	69%	-3%	-2%	5%	-9%	-7%
iwer34	81%	80%	67%	-10%	2%	18%	14%	29%
iwer35	77%	88%	49%	4%	13%	-1%	-11%	11%
iwer36	98%	91%	67%	4%	5%	1%	4%	13%
iwer37	99%	77%	64%	-5%	6%	3%	31%	19%
iwer38	83%	76%	56%	2%	4%	-7%	-1%	6%

Color coding: Green means good performance and red means poor performance

Monitoring Interviewer Performance: Data Quality Indicator

- Keystroke paradata: the record of every key stroke and measures of elapsed time(collected in many CAPI studies)
- 3 meaningful factors were identified using principle component analysis
  1. Reading question text too quickly
  2. Frequent error checks
  3. High proportion of Refused or Don’t Know responses

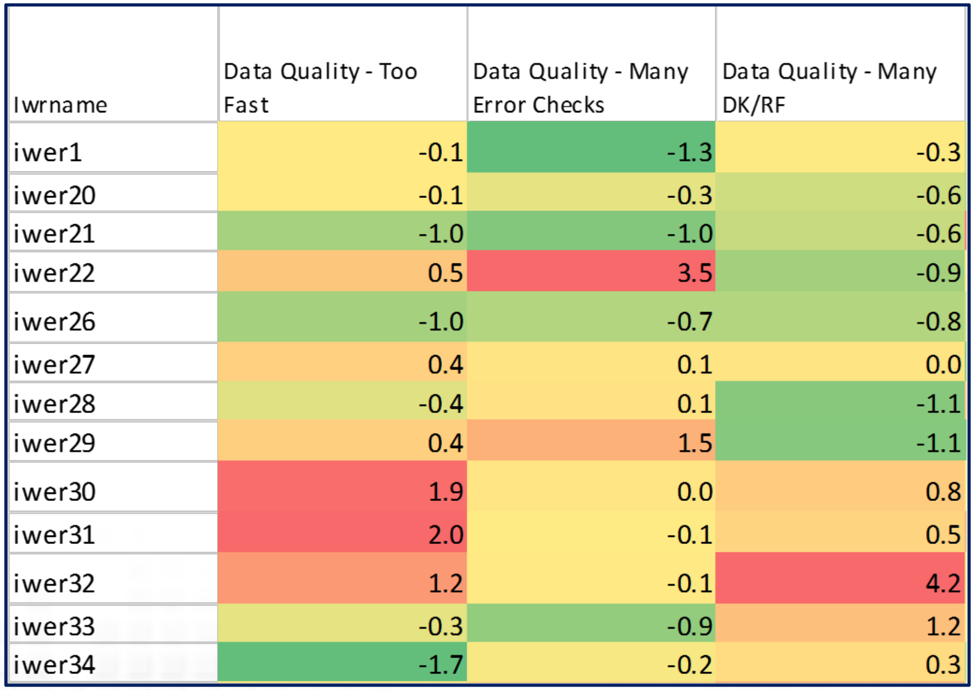
Rotated Factor Pattern (Standardized Regression Coefficients)

	Too Fast	Many Error Checks	Many DK/RF
Average Field Time per Field Visit	0.82227	-0.15392	-0.08063
Average Error Escape per Field	0.22383	0.58201	-0.07739
Average Error Suppression per Field	-0.15846	0.84359	0.06513
Average Error Jump per Field	0.43767	0.33624	0.00984
Average Back Up per Field	0.79072	0.09375	0.13965
Average Don't Know per Field	-0.06182	0.15955	0.69544
Average Refusal per Field	-0.08875	-0.16295	0.77636

Standardized regression coefficients > 0.2 are highlighted

Monitoring Interviewer Performance: Data Quality Indicator

- A standardized score (Z-score) is calculated for each factor for each interviewer as a data quality indicator.
- The Z-score indicates how many standard deviations above (red) or below (green) the interviewers' mean a raw factor score is

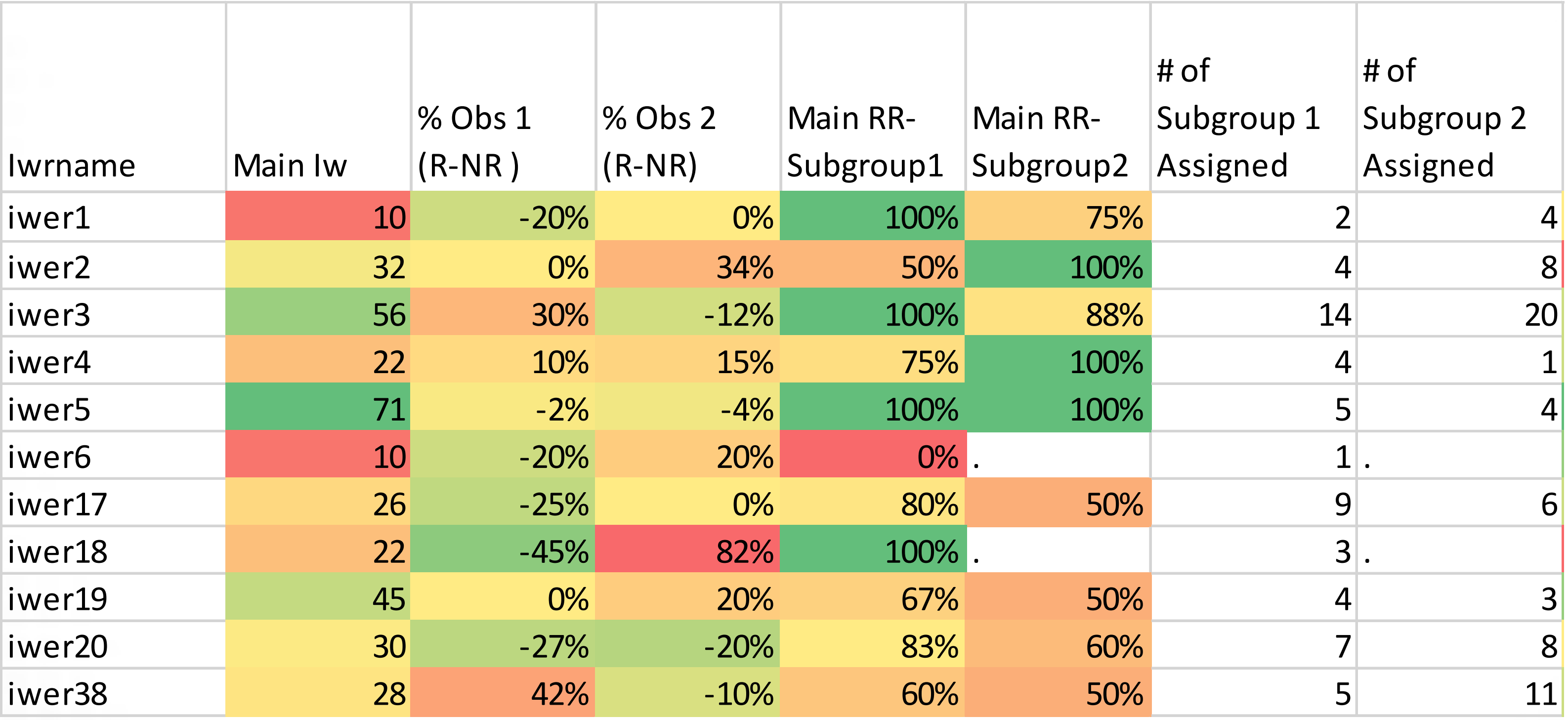


Iwrname	Data Quality - Too Fast	Data Quality - Many Error Checks	Data Quality - Many DK/RF
iwer1	-0.1	-1.3	-0.3
iwer20	-0.1	-0.3	-0.6
iwer21	-1.0	-1.0	-0.6
iwer22	0.5	3.5	-0.9
iwer26	-1.0	-0.7	-0.8
iwer27	0.4	0.1	0.0
iwer28	-0.4	0.1	-1.1
iwer29	0.4	1.5	-1.1
iwer30	1.9	0.0	0.8
iwer31	2.0	-0.1	0.5
iwer32	1.2	-0.1	4.2
iwer33	-0.3	-0.9	1.2
iwer34	-1.7	-0.2	0.3

Monitoring Interviewer Performance: Data Set Balance

- **Data set balance:** indicator utilized to minimize non-response bias
- **Data set balance metrics:**
  - % of an observed sample characteristic between Respondents(R) and Non-respondents(NR)
    - When a sample characteristic is related to key statistics of the survey, monitoring the % of the observed characteristics between R and NR helps minimize non-response error
  - Response rates for demographic subgroups

Monitoring Interviewer Performance: Data Set Balance



Iwrname	Main Iw	% Obs 1 (R-NR)	% Obs 2 (R-NR)	Main RR-Subgroup1	Main RR-Subgroup2	# of Subgroup 1 Assigned	# of Subgroup 2 Assigned
iwer1	10	-20%	0%	100%	75%	2	4
iwer2	32	0%	34%	50%	100%	4	8
iwer3	56	30%	-12%	100%	88%	14	20
iwer4	22	10%	15%	75%	100%	4	1
iwer5	71	-2%	-4%	100%	100%	5	4
iwer6	10	-20%	20%	0%	.	1	.
iwer17	26	-25%	0%	80%	50%	9	6
iwer18	22	-45%	82%	100%	.	3	.
iwer19	45	0%	20%	67%	50%	4	3
iwer20	30	-27%	-20%	83%	60%	7	8
iwer38	28	42%	-10%	60%	50%	5	11

Color coding: Green means good performance and red means poor performance

Summary

- Requires paradata collection and investment to develop the profile.
  - Once developed, can be adapted to other studies
- Uses heat maps to identify positive versus negative performance at a quick glance
- An easy tool for both methodologists and managers to identify individual areas of concern at interviewer level
  - Effort and productivity
  - Productivity after adjusting for difficulty
  - Measurement error
  - Non-response error

Current Use

- Used daily by field managers. Reviewed weekly in management team meeting.
- Recent examples to identify:
  - Interviewers with consistently high hours per case
  - Interviewers with low eligibility PAIP scores
  - Frequent error checks
  - Lower response rates with specific demographic subgroups
- Interventions:
  - Remind all interviewers of study requirements/goals
  - Discuss specific cases with identified interviewers
  - Re-training on area of concern
  - Identify "best" performing interviewers and have them share their strategies
  - Continued monitoring for improvement

Future Work

- Future development:
  - Strengths and weaknesses of each interviewer
  - Filtering the metrics by area of concern
  - Taking into account the variance of a point estimate (statistical process control)

- Possible Adaptation for New Studies:
  - Pivot table for studies with more interviewers or management levels
  - Sample assignment decisions

References

- West, Brady T. and Robert M. Groves. 2013. "The PAIP Score: A Propensity-Adjusted Interviewer Performance Indicator." *Public Opinion Quarterly*, 77(1): 352-374

- Gu, H., Couper, M.P., Kirgis,N., Parker, S. and Buageila, S. 2013. “Using Audit Trail Data for Interviewer Data Quality Management”, Presentation at the Annual Conference of the American Association of Public Opinion and Research (AAPOR)

INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER
SURVEY RESEARCH OPERATIONS
UNIVERSITY OF MICHIGAN

Thank You!

wenchang@umich.edu

hguyer@umich.edu