Implementing Adaptive Design on a Longitudinal Survey of At-Risk Youth: Empirical Evidence Based on a Deep-Dive Analysis

Hanzhi Zhou[1], Jillian Stein[1]
[1]Mathematica Policy Research, 600 Alexander Park Dr., Princeton, NJ 08540

Abstract

A number of challenges are associated with conducting longitudinal survey research with at-risk young people who are often highly mobile and difficult to engage. This paper describes our use of statistical tools for dynamic monitoring of data quality and our assessment of innovative strategies for increasing sample retention and survey completion among a sample of at-risk young adults. Drawing on multiple data sources—including baseline information and paradata—we calculate quality indicators and construct prediction models to (1) assess the representativeness of our data, (2) identify any over- or under-represented groups, (3) investigate the efficacy of our engagement and retention strategies overall and specifically for those under-represented groups, and (4) adapt our data collection efforts to maximize the representativeness of our data. The findings from this research add to the knowledge base regarding the use of alternative measures of quality in survey practice and the efficacy of using texting and social media, as tools for retaining and engaging sample members in longitudinal research.

Key Words: Representativeness, paradata, dynamic monitoring, quality indicators, survey reminders, social media, text messaging

1. Background and Motivation

YouthBuild is a national second-chance program that offers education and employment training to at-risk young people, many whom have dropped out of high school. In 2010, the Employment and Training Administration, part of the U.S. Department of Labor, contracted with MDRC, Mathematica Policy Research, and Social Policy Research to conduct a rigorous random assignment evaluation of the YouthBuild program. The primary source of data for the impact analysis are three surveys that measure outcomes related to youths' educational attainment, employment, and involvement with the criminal justice system. The surveys are administered approximately 12, 30, and 48 months after youth enrolled in the study. Given the longitudinal nature of the study, and the fact that our study members are highly mobile and from populations with historically low propensities to respond to surveys, such as young, male, nonwhite, and low-income populations, our study team recognized the need to proactively combat survey nonresponse and associated nonresponse bias (Abraham, Maitland, and Bianchi 2006). To do this our team implemented a two-pronged approach that included (1) dynamic monitoring of data quality using R-indicators, and (2) analysis of contemporary methods to engage and retain our sample across the three survey waves.

R-indicators- As noted by Groves and Heeringa (2006), one way to reduce potential nonresponse bias is to monitor the covariance between the survey analytic variables of interest (Y) and response propensities (ρ_X) in relation to the overall response rates ($\bar{\rho}$).

$$Estimated\ NR\ Bias = \frac{Cov(Y, \rho_X | X)}{\bar{\rho}}, \text{(Groves \& Heeringa, 2006)} \quad [1]$$

$$X: frame\ data; paradata$$

Accordingly, our analytic approach incorporates the use of R indicators to monitor survey quality across survey waves and across time points.

Analysis of alternate contact strategies- In addition to monitoring multiple measures of data quality to identify potential sources of nonresponse bias, our research investigates innovative approaches to engage young sample members. Expanding beyond traditional methods for locating and contacting sample members, such as sending letters and phone calls, our team incorporated texting and Facebook into our overall notification plan and conducted analysis to examine whether using more contemporary modes of communication improved our ability to locate and contact our study members. This paper investigates the efficacy of these engagement and retention strategies overall and specifically for those in under-represented groups identified through our ongoing monitoring.

In sum, our evaluation team took a two-pronged approach to combat the challenges of conducting longitudinal survey research with at-risk young adults who are highly mobile and difficult to engage. First, we employed statistical tools for dynamic monitoring of data quality across survey waves and second we incorporated the use of testing the social media into our overall engagement and retention plan across the three survey waves. This paper describes how we calculated our quality indicators and constructed prediction models to address three research objectives: (1) assess the representativeness of our data, (2) identify any over- or under-represented groups, and (3) investigate the efficacy of our engagement and retention strategies overall and specifically for those under-represented groups.

2. Methods

2.1 Data

The evaluation includes 3,436 youth ages 16 to 24 at the time of enrollment. The youth were enrolled in the study on a rolling basis, from August 2011 through January 2013. We are currently in the fifth year of this seven-year study. The surveys are offered in web and computer-assisted telephone interviewing modes with field locating for the hardest-to-reach cases. So far, we have completed the 12- and 30-month data collections and are a little more than half-way through our 48-month data collection.

2.2 Sample Characteristics

The overall study sample is young, primarily male (64 percent), and predominately black (62 percent) or Hispanic (14 percent). About 64 percent of the sample was randomly assigned to the evaluation treatment group, meaning they were to receive YouthBuild services.[1] Table 1 provides descriptive statistics for the sample.

Table 1: Descriptive Statistics of YouthBuild Sample at Baseline

Variable	Percentage or mean
Evaluation treatment group	64
Male	64
Age (average)	20.2 years
White, non-Hispanic	15
Black, non-Hispanic	62
Other, non-Hispanic	8
Hispanic	14
Has child or children	31
In stable housing	76
In temporary housing	18
Other or unknown housing	6
Parole officer	5
N	**3,436**

2.3 Research Questions and Analytic Approach

2.3.1 Phase I analysis: employing R-indicators

On one front, we used R-indicators to help gauge the extent to which respondents represent the full sample and to monitor the variability of subgroup response rates defined by a set of baseline characteristics in the model. More specifically we wanted to answer three research questions:

1. How representative are our data collected thus far (using overall R-indicators)?
2. What characteristics of the sample drive this representativeness (using variable-level partial R indicators) and are any group(s) under- or over-represented (using categorical-level partial R indicators)?
3. What do the trends of representativeness look like over the entire period of data collection for each wave and compared across the three waves (using trend plots)?

Our study will use the following indicators as proposed by Schouten, Cobben, and Bethlehem (2009):

- **Overall R-indicator:** to evaluate representativeness of the respondent population as compared to the sample population
- **Category-level unconditional partial R-indicator:** to evaluate which subgroups of a variable or a cross/interaction of variables are over- or under-represented

To build these measures of quality, we first identified our variables of interest from frame data (baseline data), including age group, gender, race/ethnicity, RA outcome (treatment/control), housing status, and whether assigned to a parole officer and used them as predictors in a logistic regression model to estimate response propensities and calculate

[1] As is often the case in random assignment evaluations involving social programs (Boruch, Weisburd, Turner, and Littell 2009), assignment to the treatment and control group varied across programs. The average treatment allocation is 60/40.

(partial) R-indicators. We then monitored the (partial) R-indicators based on logistic regression models and monitored the response rates in subgroups defined by predictors in the model. This aspect generally corresponded to monitoring the numerator of equation [1]. We also used the partial R indicators for the purpose of forming nonrespondent profiles and strata for adaptive survey designs in the final cohorts of the 30-month data collection and for all cohorts during the 48-month data collection.
We look at both point-in-time and longitudinal comparisons, as reflected in the bar charts and trend plots (Figures 1 through 5), respectively.

2.3.2 Phase II analysis: employing prediction models

In the second phase of our work we examined the efficacy of our contact strategies for retaining and engaging our sample over time. More specifically, we were interested in answering the following research questions:

1. What is the relative efficacy of contact strategies for engaging sample members?
2. What strategies are most effective for under-represented subgroup(s)?

To answer these questions, we used prediction models fitted on both time-invariant covariates (e.g. contactability variables such as whether provided a phone number, or any type of social media at baseline, whether had unlimited texting plan at baseline) and baseline demographic characteristics (e.g. age, gender, race/ethnicity, housing status in the form of dummy variables). Predictors in the final prediction model were decided through a series of model selection procedures including factor analysis and stepwise regression such that they constitute a set of strong predictors for the response propensities and show little multicollinearity. On top of these predictors, we also included survey strategies variables of primary interest, namely, locating or contacting attempts defined by type and timing of media notifications (Facebook message, text message, email, or letter reminder).

3. Results

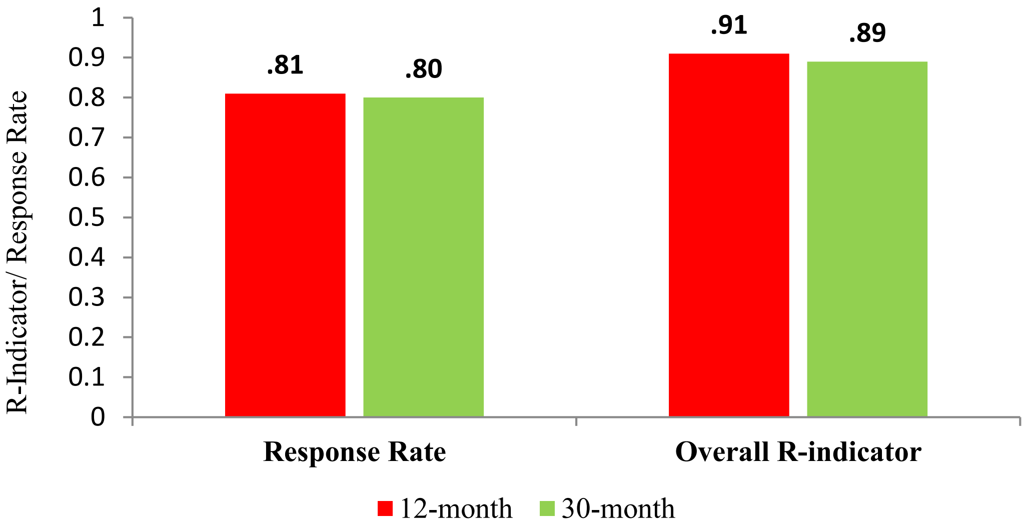
We now discuss our results in the following order: (1) static point-in-time analysis for the completed 12- and 30-month follow-ups, which helps in understanding the data quality toward the end of data collection; (2) dynamic trend analysis, which compares patterns across all three waves of data collection and helps to monitor the ongoing 48-month follow-up; and (3) analysis of the relative efficacy of contact strategies based on a static prediction model.

3.1 Static point-in-time analysis

To answer our first research question regarding the overall quality of our data, we first examined the final response rates and overall R-indicators for the two completed data collections, the 12- and 30-month follow-up surveys. In both rounds of data collection, we obtained high response rates and our data were very representative as measured by the overall R-indicator (Figure 1).

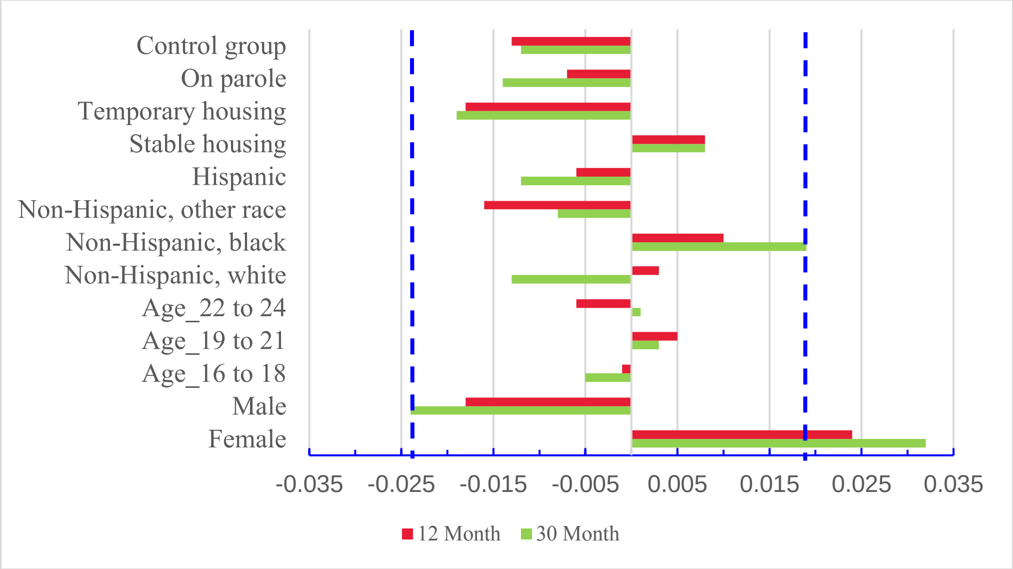
Next, to examine what groups were over- or under-represented, we looked to the partial R–indicators, which ranges from -.5 to +.5, where 0 is ideal, -.5 equates to under-representation, and +.5 equates to over-representation. We established a cutoff value of .02, which is demarcated by blue vertical lines on the slide. This cutoff value helps the study team narrow its focus on the groups most over- or under–represented, enabling the team to focus on variables that exceeded this threshold.[2] As shown in Figure 2, the only category that exceeded this threshold was gender, in which males were clearly under-represented and females were over-represented in the 12- and 30-month data (Figure 2).

Figure 1: Response Rates and Overall R-Indicators for 12- and 30-Month Data Collection



Overall R-indicator values range from 0 = Not at all representative to 1 = Perfect representation

Figure 2: Static Categorical–Level R-Indicator for 12- and 30-Month Data Collection



Categorical-level R-indicator ranged from -.5 to +.5

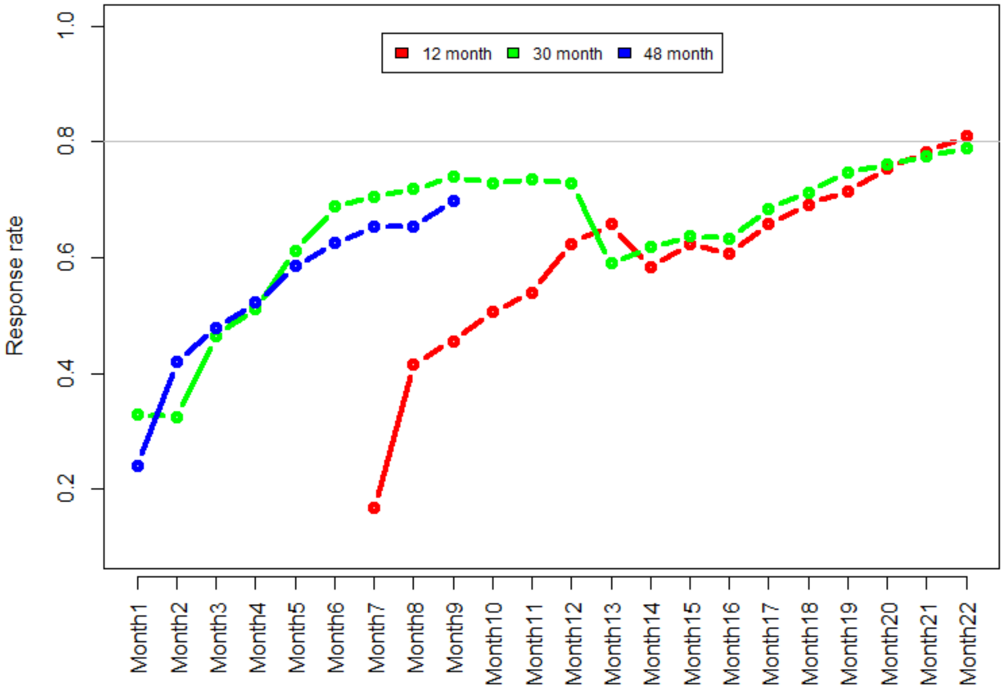
0 is ideal; -.5= under-represented; +.5 = over-represented

[2] A scan of the literature did not reveal any universal rule of thumb for choosing a threshold value for partial R-indicators. For YouthBuild data, we decide on a threshold value of 0.02 based on the relative magnitudes between categories for categorical-level partial R indicators.

3.2 Dynamic trend analysis

Because static measures do not help us assess the quality of the 48-month data so far, we show the trend plots for the three waves of data collection simultaneously, for the overall response rate and the overall R-indicator, respectively (Figures 3 & 4) as well as for the categorical partial R-indicator for the most severely under-represented group: males (Figure 5). We plot the horizontal grey line at the 80% level as our target in Figures 3 & 4 and -0.02 as our cut-off threshold for under-representativeness in Figure 5.

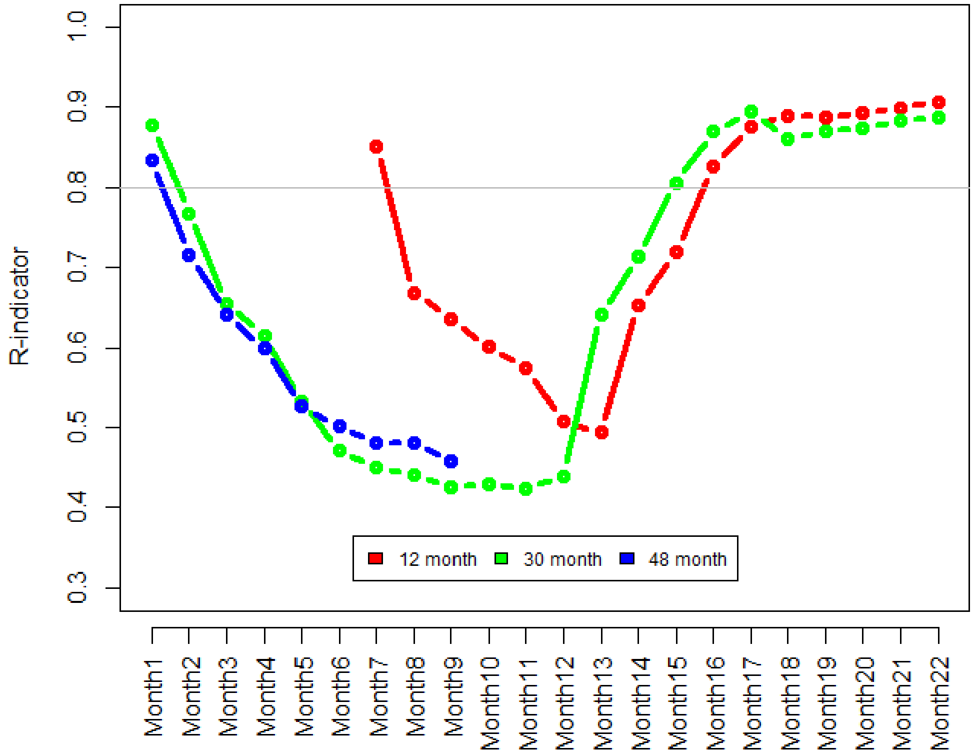
Figure 3: Response Rates for 12-, 30-, and 48-Month Data Collections



Note: The 12-month data collection was delayed due to the Office of Management and Budget clearance process; hence, the first six cohorts were released together. This is why the first time point in the 12-month survey starts later than in the other waves.

For all three waves of follow-up, the response rate generally increased from month to month, except for a sharp drop between months 12 and 13. This sharp drop between releases prompted us to examine the distributions of several demographic variables by cohort. Relative to other cohorts, the 13th cohort consists of high proportion of males, which explains the decrease in response rate. Between months 1 to 12, the R-indicator decreased steadily from around 0.9 to 0.45, as would be expected – since the R-indicator is a function of the variance of response propensities, the higher the variability of the response propensities, the less representative the data; during the data collection period, as late responders (with typically low response propensities) of the earlier released cohorts start to respond and early responders (with typically high response propensities) of just released cohorts also start to respond, we should expect to see a mixed respondent pool of increased variability in their response propensities, and as a result, a decrease in representativeness of the data collected. It then increased sharply to about 0.85 in months 13 to 17, and stayed in that range until month 22, namely the end of data collection for both 12 month and 30 month follow-ups.

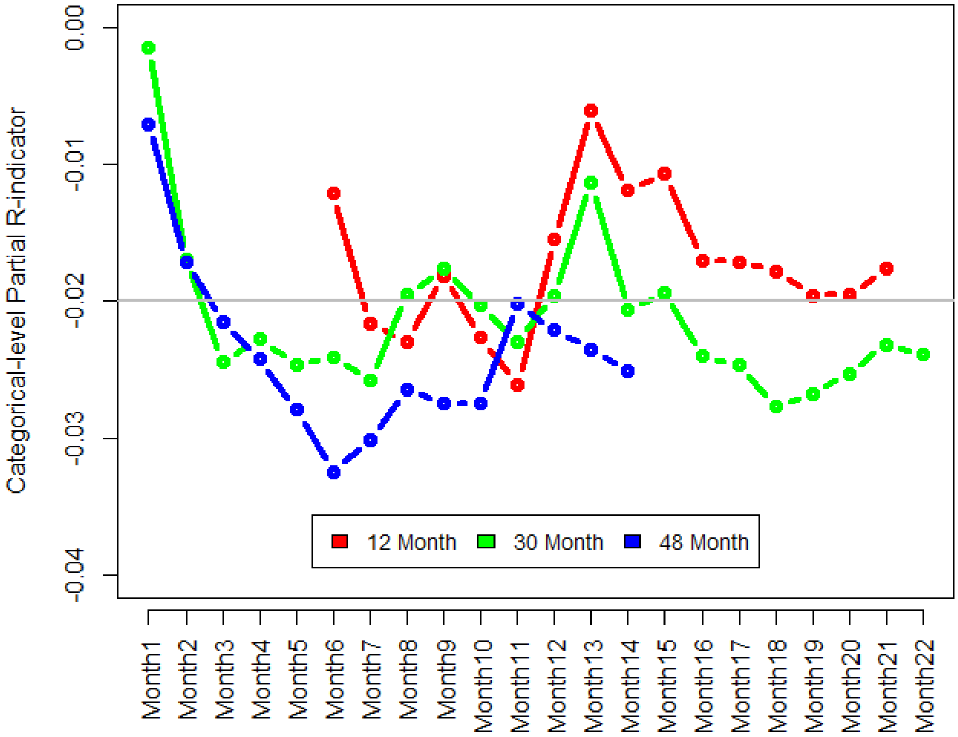
Figure 4: Overall R-Indicator for 12-, 30-, and 48-Month Data Collections



Overall, the 30-month data demonstrated a very similar trend to the 12-month data for both the response rate and the R-indicator, but the magnitudes differed before month 13: the 30-month survey yielded higher response rates yet equivalently less representative data than the 12-month survey, although this difference narrowed over time. After month 13, we saw a nearly perfect overlapping between the trend lines for 12 and 30 months. Such results were fairly informative and facilitated the monitoring of the 48-month data collection; presumably, if we did not change the follow-up strategies, we would expect to see a similar trend for the 48 month follow-up as we did for the 30 month follow-up. Tracking response rates over time enables us to see that response rates for the 48-month data collection are lagging slightly but the overall representativeness of our 48-month data looks strong relative to this time in the 12 and 30 month data collections.

Next, because we had identified males as the only under-represented group in the 12- and 30-month data, we plotted the partial R-indicator value for males across the three waves (Figure 5). As illustrated in Figure 5, males continued to be under-represented in the 48-month data collection and lagged slightly behind previous rounds, making them a target population for additional analysis in the second phase of our research.

Figure 5: Trend of Categorical Partial R-Indicators at 12-, 30-, and 48-Month Data Collections for Males



It is also worth noting that in an impact evaluation study context using a randomized experimental design, it is important to look at the separate trends for the treatment group and the control group in terms of data quality measures. Any unexpected differences in survey completion patterns between the two subgroups at the data collection stage could have a non-ignorable effect on future data analysis for causal inference. For all three waves of data collection, we do not see drastic differences between the treatment and the control for the Youthbuild sample so far and do not show figures here.

3.3 Efficacy of sample retention and engagement strategies

To answer our third research question regarding the relative efficacy of our four main notification strategies (Facebook messages, text messages, emails, and letter reminders), we used logistic regression, controlling for baseline measures of gender, race and ethnicity, age, housing status, number of social media accounts, whether they provided us a phone number or address, whether they had unlimited texting at baseline, and other notifications received during 30-month data collection. As illustrated in Figure 6, Facebook and texting were associated with greater odds of completion than letters or emails among the full sample. These results aligned with our hypothesis that conventional modes of communication would be less effective than more contemporary modes of communication among this young adult population.

Next we examined these results for our subgroup of interest: males. As illustrated in Figure 7, their story resembled the full sample, with text and Facebook messages positively related with odds of completion. Among the male-only sample, text messaging seemed slightly more effective than for the overall sample, whereas Facebook messages were slightly less effective for males than for the overall sample.

Figure 6: Adjusted Odds of Completing the 30-Month Survey for the Full Sample

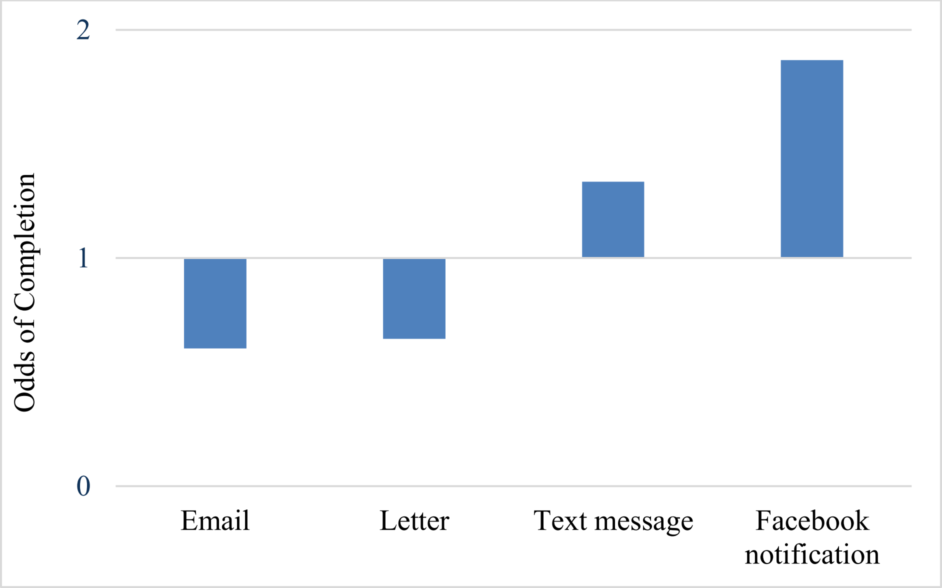
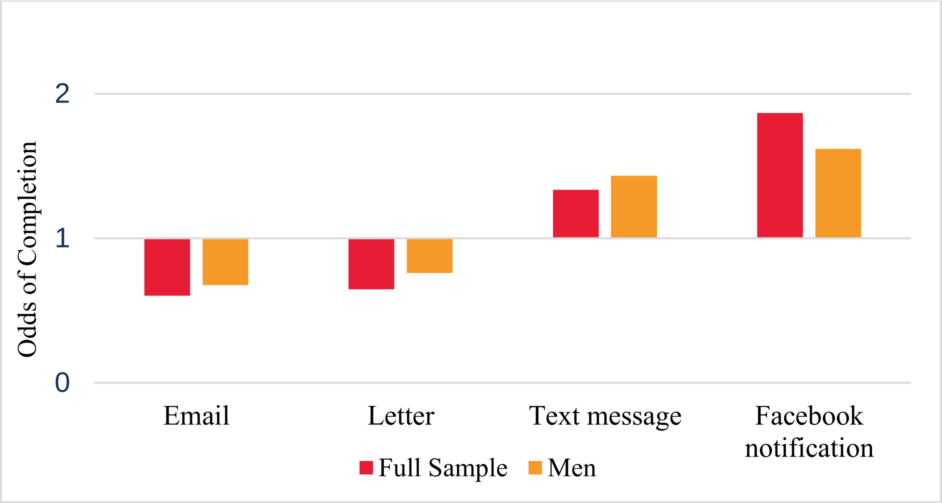


Figure 7: Adjusted Odds of Completing the 30-Month Survey for the Full Sample and Men



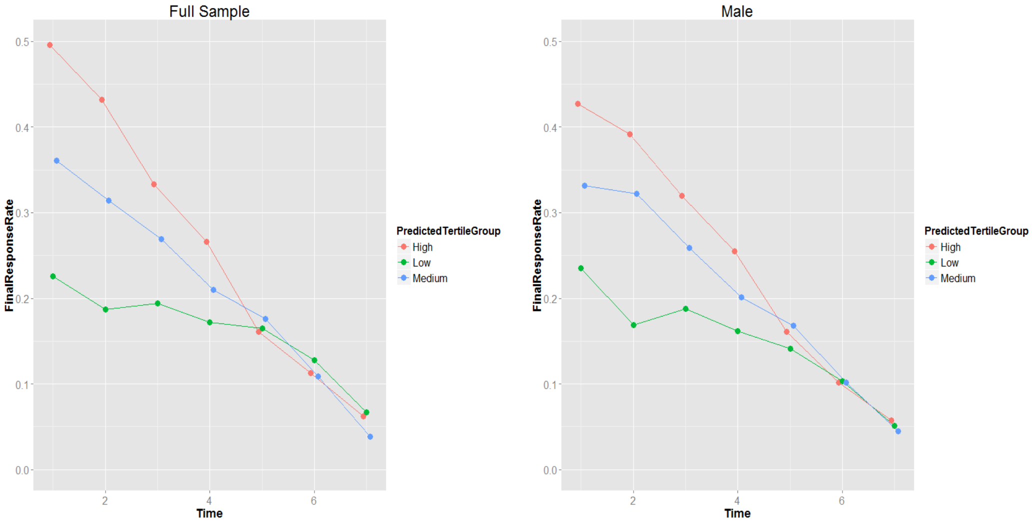
4. Discussion

Two sets of research objectives comprised our analysis: (1) assessment of the representativeness of our data using multiple metrics of quality and (2) investigation of survey notification strategies most related to survey completion overall and for any under-represented groups. Through dynamic monitoring of R-indicators, our analyses identified males as under-represented. An assessment of the efficacy of our contact strategies for staying in touch with and notifying sample members that their survey was available, found that Facebook and texting were associated with greater odds of completion compared to letters or emails. These results align with our original hypothesis that conventional modes of communication might be less effective for this population. When looking at whether these strategies were particularly effective for males, our under-represented group, our analyses showed that none of our existing notification strategies were especially effective for men. This led us to ask what we could do differently to achieve higher male completion rates.

To begin to try to answer this question, the evaluation team conducted two sets of qualitative interviews. The first was an informal conversation with young men similar in age to our sample members in which we showed them our text message and Facebook messages and asked for their feedback. We held the second set of conversations with staff at three YouthBuild programs located in three states focusing on the methods they found most effective when trying to contact their young male program participants. Feedback from these interviews led us to draw two main conclusions: (1) message content that is shorter, less formal, and personalized was more appealing to males than our current messages; and (2) finding incentives other than money that appeal to male sample members might help grab sample members' attention and encourage them to complete our survey. The study team is currently planning adaptations to the data collection notifications to incorporate this feedback.

Another direction we could pursue is combined use of partial R-indicators and *dynamic* prediction models to build profiles of the active cases (e.g. males that haven't completed the survey) to whom we might target our resources and implement tailored interventions. This prediction model expands upon the *static* prediction model we used previously, with more auxiliary variables (e.g. time-varying paradata) included as additional predictors. In fact, we conducted a preliminary analysis based on this idea, where we retrospectively validated the dynamic prediction model through fitting them at seven different time series data points in the later phase of our 30-month follow-up, following the method of Wagner and Hubbard (2014) to determine the accuracy of our profiling of the active cases for both the full sample and male subsample. The active cases were divided into high, medium and low tertile groups based on their predicted response propensities and were compared with the final response rate. The models appear promising for predicting survey completion in the early and middle stages of data collection; as reflected in Figure 8: for both the full sample and male subsample, the three lines don't cross one another except for at the very end of data collection when nonresponders' behaviors aren't accurately predicted by the model.

Figure 8: Retrospective Validation of Model Estimation



Our results make intuitive sense, but we would like to close by highlighting some of the limitations to our analyses, which present opportunities for future research that could build on this work. First, our analyses relied on observational data, but research using experimental designs is needed. While the evaluation team has conducted experimental research to evaluate texting and was able to conclude that texting significantly increased odds of survey completion compared to a control group whom did not receive text messages (Skaff, Stein, and Hurwitz, 2016), more research is needed to examine the relative benefits of specific interventions compared with other interventions. Furthermore, incorporating cost data would add a very important layer of information that could inform our assessment of the relative cost-efficacy of particular contact strategies relative to others. A final limitation of our research, were design elements that made it hard to adapt during data collection, namely overlapping survey rounds and the rolling and small sample releases, making it operationally hard to target males systematically.

5. Conclusions and Next Steps

To conclude, our work contributes to the survey research and operations literature in three general areas: **(1)** ***data collection monitoring-*** our statistical analysis using the YouthBuild data demonstrated the utility of static and dynamic monitoring of R-indicators during data collection. On one front, we showed that static R-indicators are useful in identifying subgroups for special monitoring and treatments. On another front, we demonstrated how trend analysis of R-indicators is useful when managing active data collections because they make it possible to examine where one is in the current data collection relative to where one was at a similar time point in previous rounds. Furthermore, our experience taught us that it is easier to develop trend plots of quality indicators if you take snapshots of the data at regular intervals, in our case, monthly, to compare patterns across waves of data collection. **(2)** ***assessment of contemporary survey notification strategies-*** our research demonstrated that innovative contact and retention strategies, namely Facebook and texting strategies are associated with greater response compared with traditional methods (letters and emails). Although none were especially effective in drawing in males, this prompted us to investigate into qualitative interviews which suggest that our team can adapt our survey notifications so they are more appealing to males (for example by making the messages shorter, personalized, and more informal or by sending incentives other than cash); these preliminary findings should be further investigated and tested with more rigorous experiments for adaptive design. ***(3) adaptation of data collection-*** While our team has not implemented many adaptations, largely because our current strategies have led to highly representative data for the 12 and 30 month follow up surveys, our research has illuminated some potential directions for adaptation in the future. Beyond modifying our messaging for male respondents, future research could use dynamic predictive models to guide resource allocation to the most important cases with the highest propensity to respond. For example, if we couldn't afford to send all cases to the field we could prioritize cases with a higher propensity to respond. Our preliminary results about the paradata-fed models validated their relative accuracy but future research would be needed to test out their implementation in responsive or adaptive designs. During the process, there is room to develop better models through leveraging more data with more sophisticated estimation methods (for example, Bayesian models that incorporate prior information available from both previous waves and ongoing data collection, while allowing for great flexibility in the specification of prior distribution on the regression coefficients.) potentially improving the prediction of the hardest-to-reach cases towards the end of data collection.

Acknowledgements

The authors would like to thank the U.S. Department of Labor, Employment and Training Administration, MDRC, and Social Policy Research Associates for funding the project. The authors would also like to thank their colleagues at Mathematica: Dan O'Connor, Danni Tu, and the YouthBuild team for their support in the work and their helpful comments on the paper.

References

Abraham, K. G., A. Maitland, and S. M. Bianchi. "Nonresponse in the American Time Use Survey: Who Is Missing from the Data and How Much Does It Matter?" *Public Opinion Quarterly,* vol. 70, 2006, pp. 676–703.

Boruch, R., D. Weisburd, H. Turner, and J. H. Littell. "Randomized controlled trials for evaluation and planning." In *Handbook of Applied Social Research Methods* (2nd ed.), edited by L. Bickman and D. Rog. Thousand Oaks, CA: Sage Publications, 2009, pp. 147–181.

Goble, L., Stein, J., & Schwartz, L. K. Approaches to increase survey participation and data quality in an at-risk youth population. Presented at the FedCASIC Conference, Washington, DC, March 2014.

Groves, R. M., & Heeringa, S. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. Journal of the Royal Statistical Society Series A: Statistics in Society, 169, 439–457.

Little, R. J., & Vartivarian, S. L. (2005). Does weighting for nonresponse increase the variance of survey means? Survey Methodology, 31, 161–168.

Lundquist, P., & Särndal, C. E. (2013). Aspects of responsive design with applications to the Swedish Living Conditions Survey. Journal of Official Statistics, 29(4), 557–582.

Särndal, C. E. (2011). The 2010 Morris Hansen Lecture: Dealing with survey nonresponse in data collection, in estimation. Journal of Official Statistics, 27(1), 1–21.

Shouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. Survey Methodology, 35, 101-113.

Shouten, B., Shlomo, N., & Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. Journal of Official Statistics, 27, 231–253.

Skaff, A., Schwartz, L.K, & Stein, J. Harnessing social media in survey research. Paper presented at the American Association for Public Opinion Research Conference, Anaheim, California, May 16, 2014.

Stein, J., Skaff, A., Schwartz, L.K., and O'Connor, D. "Social Media in Survey Research: Can Facebook Friendship Enhance Traditional Survey Strategies?" Presented at the FedCASIC Conference, Washington, DC, March 19, 2014.

Wagner, J. (2012). A comparison of alternative indicators for the risk of nonresponse bias. Public Opinion Quarterly. doi: 10.1093/poq/nfs032

Wagner, J., & Hubbard, F. (2014). Producing unbiased estimates of propensity models during data collection. Journal of Survey Statistics and Methodology, *2*(3), 323–342. doi: 10.1093/jssam/smu009