Received: 21 July 2020 | Accepted: 19 August 2021

DOI: 10.1111/rssa.12750

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

ORIGINAL ARTICLE

Transitioning a panel survey from in-person to predominantly web data collection: Results and lessons learned

Paul P. Biemer[1,2] | **Kathleen Mullan Harris**[2] | **Brian J. Burke**[1] | **Dan Liao**[1] | **Carolyn Tucker Halpern**[2]

[1]RTI International, Research Triangle Park, USA

[2]The University of North Carolina at Chapel Hill, Chapel Hill, USA

Correspondence
Paul P. Biemer, RTI International, 3040 E. Cornwallis Road, Research Triangle Park, North Carolina 27709, USA.
Email: ppb@rti.org

Abstract

Over the last two decades, in-person interviewing costs continued to increase while the data quality advantages traditionally identified with this data collection mode have faded. Consequently, some longitudinal surveys have begun transitioning from in-person to web data collection despite risks to data quality and longitudinal comparability. This paper addresses the major issues involved in the transition process and proposes a multi-sample, multi-phase responsive design that attempts to minimize the data quality risks while preserving the considerable cost savings promised by the transition. The paper describes the design as it was applied to the National Longitudinal Study of Adolescent to Adult Health (Add Health)—a nationally representative panel survey of around 20,000 adolescents selected from grades 7 to 12 (typically 13 to 18 years of age) in the 1994–95 school year. Also described are key results from several experiments embedded within the design and the analysis of mode effects. Also presented are some lessons learned and recommendations for other

in-person panel surveys that may be contemplating a similar transition to web or mixed-mode data collection.

KEYWORDS

incentive experiments, mode effects, mode transition, modular questionnaire, non-response follow-up, responsive design

1 | INTRODUCTION

In-person field interviewing has long been considered the gold standard for collecting panel survey data because it generally provides higher response rates than other interview modes (Groves & Couper, 1998; de Leeuw & van der Zouwen, 1988; Watson & Wooden, 2009). For collecting data on complex phenomena, in-person surveys can provide more accurate and complete responses than self-administered surveys because concepts can be clarified by interviewers who can also motivate respondents to more fully complete the questionnaire (Heerwegh, 2009; de Leeuw & Collins, 1997; de Leeuw & van der Zouwen, 1988). Although social desirability bias can be an issue for sensitive topics when an interviewer is present, computer-assisted self-interviewing (CASI) can be embedded in the in-person interview to address that concern (Brown et al., 2013; O'Reilly et al., 1994). The major disadvantage is cost, which can be many times greater than self-administration using a combination of web and paper (web/paper) questionnaires (Dillman, 2017; Dillman et al., 2009; Villar & Fitzgerald, 2017). Contributing to high costs are the declining response rates in household surveys because greater field effort is required to contact and interview respondents than ever before (Bianchi et al., 2017; de Leeuw & de Heer, 2002). Data timeliness can also be adversely affected by the additional fieldwork effort because the data collection period is then often protracted (Starick & Steel, 2012). There is also some evidence that data quality for in-person surveys may not be superior to a well-designed web or mixed-mode survey for many topics (Heerwegh et al., 2007). Interviewer effects can be an important reason for this (Biemer & Lyberg, 2003). Given less concern about reducing data quality relative to in-person surveys, it is not surprising that an increasing number of ongoing in-person surveys are transitioning to a self-administered mode to reduce costs (Biemer et al., 2018; Murphy et al., 2018). Still, transitioning from in-person to self-administration must proceed cautiously to reduce the risks of data quality deterioration due to non-sampling errors (Sakshaug et al., 2010). This is particularly true for estimates of longitudinal change in panel surveys where even random or unsystematic errors can be severely biasing (Abowd & Zellner, 1985; Cernat, 2015a, 2015b; Dillman, 2009).

The data collected for this paper are from the National Longitudinal Study of Adolescent to Adult Health (Add Health) (Harris et al., 2019). In Wave V of this panel survey, we implemented a unique survey design and series of experiments for transitioning the survey from in-person interviewing to self-administration (via the web, primarily) with an in-person follow-up component. As shown in the paper, this design substantially reduced costs, improved data quality and provided for estimation of and adjustment for mode effects in cross-sectional and longitudinal analyses. Section 2 provides a brief review of the literature on transitioning from interviewer (or in-person) to self-administration. Section 3 describes the design that was implemented in Add Health Wave V, whereas Section 4 recounts some of the experiments and special features embedded in the design. Section 5 summarizes the mode-effect analysis that was conducted after data collection. Finally, Section 6 discusses some lessons learned from the Add Health Wave V

experience, concluding with a discussion of the implications of the findings for the future of Add Health and other longitudinal surveys that transition from in-person to a self-administered mode of data collection.

2 | TRENDS FROM IN-PERSON ADMINISTRATION TO SELF-ADMINISTRATION

As data quality and cost concerns continue to mount for them, more and more interviewer-administered surveys are transitioning to web or mixed-mode data collection—especially combinations of web, paper and face-to-face questionnaire administration (see, e.g. Olsen et al., 2021). For general household surveys, the move from interviewer-administered to web is confronted by numerous design issues related to (a) survey frame construction and coverage; (b) household and within-household selection; (c) questionnaire design and length; (d) survey error, especially nonresponse; and (e) the risk of methodological effects caused by the transition that could bias comparisons with pre-transition estimates. The American Association for Public Opinion Research *Task Force Report on Transitions from Telephone Surveys to Self-Administered and Mixed-Mode Surveys* (Olsen et al., 2021) discusses these and other issues that apply to most interviewer-administered surveys including panel surveys.

Although there is some overlap, the issues for transitioning an ongoing longitudinal survey from interviewer-administered to web can be very different from those for cross-sectional surveys. For instance, for an ongoing panel survey, issues (a) and (b) mentioned earlier are of less concern because the sample has already been selected, and the within-household informant may have already been identified in prior survey waves. However, issues (c), (d) and (e) are still quite cogent, with the last issue being the chief concern in longitudinal surveys because estimating longitudinal change is usually the prime objective.

In Add Health Wave V, the sample members were quite knowledgeable about the survey, having been interviewed in person four times previously. This provided benefits and challenges for the redesign. Some of the most important challenges the redesign faced were the following:

- Tracing sample members who had not been contacted since Wave IV—a time lag of more than 8 years—even longer for sample members who were not contacted at Wave IV.
- Convincing sample members who were quite familiar with the in-person interviewing approach used for prior waves to complete the survey online with no interviewer interaction.
- Appropriately incentivizing sample members to complete a relatively lengthy web questionnaire originally designed for in-person administration.
- Redesigning the in-person questionnaire to be compatible with web/paper self-administration (e.g. by reducing the interview from 90 to 50 min) without introducing measurement errors and other mode effects that could threaten comparability with prior Add Health waves.

These challenges are not unlike those facing any panel survey attempting to transition from in-person to mixed-mode data collection. Fortunately, our efforts could capitalize on the experiences of other surveys reported in the literature that successfully carried out similar transitions. An important example is the Innovation Panel (IP) (Institute for Social and Economic Research, 2020), a household panel survey that exists for the purpose of methodological testing and development in the context of Understanding Society: The UK Household Longitudinal Study (University of Essex, Institute for Social & Economic Research, 2019). Now in its 11th wave, the

IP began its transition from in-person to web in its 5th wave. To increase response rates, the IP incorporates in-person follow-up of all web non-respondents. Another survey that recently transitioned to the web in combination with mail/paper and in-person collection is the annual Health and Retirement Study (HRS). In the most recent survey administration, HRS used web collection for the one-half sample its staff previously had interviewed by phone and continued with in-person collection for the other one-half sample (HRS, 2019). These modes are then switched for each one-half sample in the next wave of interviews. The Panel Study of Income Dynamics (PSID) also uses the web during the off-core collection years. In 2014, the PSID used web/paper to collect information about childhood experiences (McGonagle et al., 2012), and in 2016, it was used to collect information on a range of personal topics (Freedman, 2017). The Canadian Labour Force Survey uses interviewer-assisted recruitment for the first wave of data collection; however, it transitioned to web/paper for all remaining waves in 2015 (Francis & Laflamme, 2015).

There is ample evidence from the literature that the total survey error can vary by data collection mode, giving rise to the so-called mode effects (see, e.g. Olsen et al., 2021). The evaluations of mode effects may involve the use of (a) 'gold standard' or administrative data record systems; (b) parallel surveys, conducted in different modes with different respondents, sometimes called 'ringfencing' or 'benchmarking'; (c) contemporaneous reinterviews of respondents on the same topics, possibly varying the mode; and (d) statistical modelling and analysis that may involve a combination of these approaches. Response rates are generally lower for web than in-person surveys (Biemer et al., 2018; Dillman et al., 2014; de Leeuw, 2018). However, with in-person follow-up of non-respondents, such dual-mode response rates can be comparable with single-mode in-person interviewing (see, e.g. Institute for Social and Economic Research, 2020). The mode analysis reported in Section 5 for Add Health demonstrates how methods (b), (c) and (d) were jointly used to evaluate mode effects.

Non-response, frame coverage and measurement errors often jointly conspire to bias longitudinal change estimates, particularly those involving waves before and after the transition (Biemer et al., 2021; Cernat, 2015b; Sakshaug et al., 2010). For example, the previous literature consistently finds that social desirability bias, primacy/recency effects and interviewer effects are important sources of differential bias when switching from in-person to self-administration transitions (Cernat, 2015a, 2015b; Cernat et al., 2016; Heerwegh, 2009; Pickery & Loosveldt, 2000). Regarding Add Health Wave V, this was very much the case as reported in the study by Biemer et al. (2021), which is discussed further in Section 5.

The next section describes the design for this transition from an in-person to a mixed-mode survey, its objectives and a few special features that were employed. Section 4 presents some key results from the experiments embedded in data collection, and Section 5 summarizes the results of the mode analysis. In Section 6, the paper concludes with a summary of the major findings, some lessons learned and implications for other panel surveys contemplating a major redesign.

3 | THE TRANSITION DESIGN

3.1 | Description of add health

Add Health is a longitudinal study of a nationally representative sample of approximately 20,000 adolescents in grades 7–12 (typically in the age range 13–18) in the United States during the 1994–95 school year. Since then, there have been four additional waves of surveys: Wave II in 1996, Wave III from 2001 to 2002, Wave IV from 2008 to 2009 and Wave V from 2016 to 2018 (see

Harris et al., 2019). The Wave V survey collected social, environmental, behavioural and biological data that are used to track the emergence of chronic disease as the cohort moves through their 4th decade of life. Add Health combines longitudinal survey data on respondents' social, economic, psychological and physical well-being with contextual data on the family, neighbourhood, community, school, friendships, peer groups and romantic relationships, providing unique opportunities to study how social environments and behaviours in adolescence are linked to health and achievement outcomes in adulthood (additional information at https://www.cpc.unc.edu/projects/addhealth). The data collection mode for Waves I through IV was in-person interviewing exclusively, whereas, as described in the next section, Wave V primarily used web/paper self-administration. In-person interviewing was also used for the follow-up of web/paper non-response and for a small benchmark sample conducted in parallel with the main survey.

3.2 | The wave V responsive design

Wave V employed a responsive design approach (Groves & Heeringa, 2006) consisting of multiple data collection phases. First, the eligible sample consisting of 19,828 persons was divided into four random subsamples—referred to as Samples 1, 2a, 2b and 3. All four samples are nationally representative and were fielded in slightly overlapping, sequential time periods spread over approximately 3 years (see Figure 1). Samples 1, 2a and 3 (whose union will be referred to as Sample MM for 'mixed mode') were conducted by a mixture of web, paper and in-person (i.e. face-to-face interviewing and CASI) data collection modes. Sample 2b was conducted in parallel with two of the mixed-mode samples—Samples 2a and 3—using the same data collection procedures used in prior waves (i.e. in-person interviewing). Thus, Sample 2b served as a benchmark sample for assessing mode effects (discussed further in Section 5). Two additional data collection phases were embedded in each mixed-mode sample. For these samples, Phase 1 employed the Choice +web/paper protocol (Biemer et al., 2018) that was then followed in Phase 2 by a non-response follow-up (NRFU) of a sample of Phase 1 non-respondents using in-person interviewing. To leverage the availability of the field interviewers and increase cost efficiency, Phase 2 (Samples 2a and 3) and Sample 2b were fielded together.

The phased structure of the sample has several advantages. First, sequentially fielding the mixed-mode samples allows for methodological experiments embedded in the earlier phases to inform and improve the design of subsequent field samples. The sizes of the samples were chosen

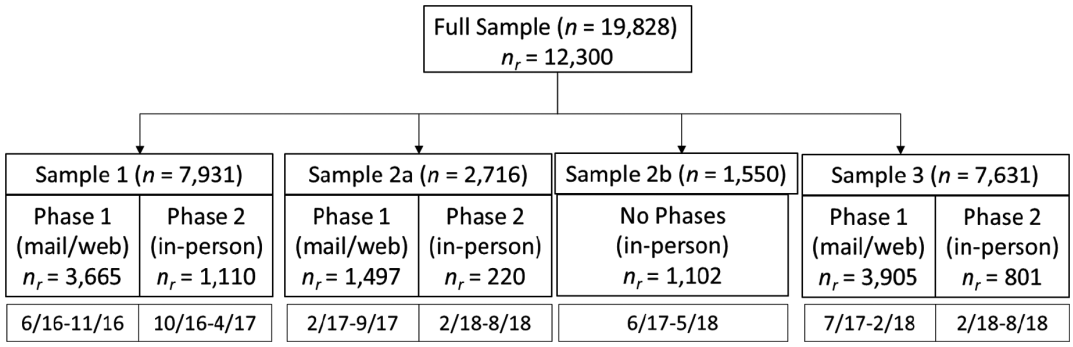

FIGURE 1 Add Health Wave V multi-sample, two-phase responsive design
Note: n_r denotes the number of respondents. The date format is MM/YY.

to achieve a levelling of resources over the duration of the project, which was advantageous for project management and cost control. Second, as shown in Section 4, the Phase 2 NRFU implemented for the mixed-mode samples elevated the overall response rates and substantially reduced non-response bias in the adjusted estimates. Third, Phase 2 also allowed for some control of costs and errors through simply adjusting the Phase 2 sampling rate. Finally, as shown in Section 5, the parallel in-person sample (Sample 2b) allowed for the evaluation of mode effects in the estimates, which were quite substantial.

The multi-sample, multi-phase benchmarked responsive design employed in Wave V is quite novel in the survey literature. Although the idea of benchmarking has been used in other panel surveys (e.g. Understanding Society: The UK Household Longitudinal Study; see Lynn, 2009), the combination of features implemented in Wave V for controlling non-response bias and measurement effects appears unique, particularly because it was the first time in Add Health's 26-year history that web/paper data collection was used in a major way. In fact, as shown in Figure 1, traditional in-person interviewing was used only for about 26% of the Wave V sample.

4 | EMBEDDED EXPERIMENTS

To pre-test the data collection systems prior to the main data collection, a small pilot study was conducted for a convenience sample of about 300 sample members in three states. One important indication from the pilot was that the response rate to the web questionnaire could be substantially lower than expected. To address this issue, we consider the options for a 'push to web' design (see Dillman, 2017, for a comprehensive review of such designs). 'Push to web' designs provide the option of responding by paper but encourage web response in some way. Ultimately, a variation of the so-called Choice +protocol (Biemer et al., 2018) was adopted. Under this design, mail and web options are offered in the first questionnaire mailing; however, to promote the choice of web, a substantial monetary incentive is promised to respondents for completing the questionnaire by web. The pilot also suggested that email prompts could be an effective non-response conversion strategy; thus, that was added to the postal prompts for sample members with known email addresses. The remaining sections describe the experiments that were embedded in Samples 1, 2a and 3.

4.1 | Sample 1 experiments

Sample 1 was partitioned into four equal random subsamples and assigned to four treatments arrayed as a 2×2 factorial design. Factor 1 varied the design of the questionnaire in two conditions: (a) the modular questionnaire featuring two separate modules of approximately equal length that were administered in tandem (i.e. Module 2 was accessible only after the respondent had completed Module 1); and (b) the content-equivalent, singular questionnaire design. For (a), each module was about 25 min in duration, and for (b), the two modules were essentially concatenated to create an approximately 50-min questionnaire. The hypothesis was that the modular questionnaire, by minimizing the initial impact of the interview length and response burden on the sample members, would lead to higher completion levels (unit and item response) and better data quality. Although the literature is somewhat equivocal and primarily based on experience in cross-sectional designs, several papers support the hypothesis of improved response using a

modular questionnaire design (see, e.g. Crawford et al., 2001; Dillman, 2007; Galesic & Bosnjak, 2009; Gummer & Roßmann, 2015; Loosveldt & Beullens, 2013).

Factor 2 compared two alternative incentive assignment protocols: (a) an experimental approach incorporating model-directed incentive assignment and (b) the standard incentive assignment that provided the same incentive to all sample members. Treatment (a) used a model to classify sample members into low- and high-response propensity groups. The former group was offered a higher value incentive package than the latter group. The classification model (see Biemer et al., 2020, for details) uses paradata from Waves III and IV to identify sample members who did not respond to either or both waves or who responded but required a much greater follow-up effort. Other variables in the model include the sample member's age, gender, race/ethnicity, state, region, an urban/rural indicator and the likelihood that the sample member can access the internet. The latter variable was based on census block-level data on internet access from the most recent American Community Survey as well as prior wave data on email account ownership. The 'internet access' variable was included because one goal of the design was to maximize the number of respondents opting to respond by web versus paper. Thus, the control group for the experiment received the singular questionnaire and the standard incentive.

Consistent with the Choice +protocol, the option of responding by web or paper was initially offered to all sample members regardless of treatment group. Response by the web was rewarded with a bonus monetary incentive of $20. All sample members received a pre-notice letter followed by the prepaid incentive that was delivered in the same mailing as the initial questionnaire. The four treatment conditions and their incentive structures are shown in Table 1.

The total values of these incentive offerings are consistent with those used in prior waves of the survey. For example, in Wave IV, the incentive began at $40 but was increased to $100 in the latter stages of the fieldwork (Aragon-Logan et al., 2010).

The results of the experiment appear in Table 2 with Factor 1 (singular vs. modular questionnaire) on the rows and Factor 2 (standard vs. model-directed incentive structure) on the columns. The overall response rate for Phase 1 is 46.6%, as shown on the bottom-right corner of the table. The three response rates shown in each cell correspond to the low- and high-propensity groups separately in regular font and the combined response rate in bold font. The following summarizes some of the key results from the table. First, we focus on the combined (bold) response rates.

TABLE 1 Incentive structure for the four treatment conditions

	Factor 2—Incentive Structure		
		Model Directed	
Factor 1—Questionnaire Design	**Standard Total Incentive = $65**	**Low-Propensity Total Incentive = $75**	**High-Propensity Total Incentive = $55**
Singular	$10 prepaid $45 promised $20 bonus if via web	$10 prepaid $45 promised $20 bonus if via web	$10 prepaid $25 promised $20 bonus if via web
Modular	$5 prepaid $20 promised: Module 1 $20 promised: Module 2 $20 bonus if both modules via web	$5 prepaid $25 promised: Module 1 $25 promised: Module 2 $20 bonus if both modules via web	$5 prepaid $15 promised: Module 1 $15 promised: Module 2 $20 bonus if both modules via web

Note: The control group cell is shaded.

TABLE 2 Sample 1 phase 1 response rates, by treatment group (Unweighted)

	Factor 2—Incentive Structure				Total
	Standard		Model Directed		
Factor 1—Questionnaire Design	Low-Propensity ($65 Incentive)	High-Propensity ($65 Incentive)	Low-Propensity ($75 Incentive)	High-Propensity ($55 Incentive)	
Singular	**48.6**		**46.8**		**47.7**
	31.5	60.3	32.0	56.9	
Modular	**45.0**		**45.8**		**45.4**
	28.0	56.8	30.4	56.3	
Total	**46.8**		**46.3**		**46.6**
	29.7	58.5	31.2	56.6	

Note: The control group cells are shaded.

The bolded response rates combine the (unbolded) propensity group rates for each incentive structure.

- The control treatment, which combined the singular questionnaire and standard incentive ($65 in total), achieved the highest response rate of 48.6%.
- The combination with the lowest response rate (45.0%) was the standard incentive package with the modular questionnaire.
- The effect of the model-directed incentive is not significant for either questionnaire version, although it appears it may have suppressed the response rate for the singular questionnaire.

Next, we focus on comparisons between propensity groups (non-bold response rates).

- The response propensity model worked well to identify the low- and high-responding sample members; however, the interventions to increase response rates for the low-propensity cases were not successful.
- For the low-propensity groups under both questionnaire versions, the model-directed incentive increased response rates slightly but not significantly so.
- For the high-propensity group, there was no significant difference between the incentive treatment response rates for the modular questionnaire. For the singular questionnaire, response rates for the model-directed treatment were significantly lower than those for the standard incentive but only at the 10% level.

The results indicate that the modular questionnaire did not elicit a higher unit response rate. Apparently, the burden of completing a single long questionnaire was not an important factor for Add Health respondents. The response rate for Module 1 of the modular questionnaire was essentially the same as the response rate for the singular questionnaire. However, only 95% of those completing Module 1 also completed Module 2; thus, the response rate for the whole questionnaire (both modules) fell short of that for the singular version. As reported in Powell et al. (2018), item non-response rates (which averaged about 4.1%) and breakoffs (which averaged 1.3%) did not vary significantly across treatments.

These results seem to confirm Gummer and Roßmann's (2015) finding that the effect of interview length on survey completion is moderated to a large extent by the characteristics of the sample members, especially in a longitudinal survey context. Sample members who are more motivated to

participate, experienced with survey taking, knowledgeable about the survey's content and familiar with the survey sponsor tend to be less affected by interview length or duration. These characteristics are shared by most Wave V sample members who have participated in anywhere from one to four prior waves, which may explain the null effect. Based on the Sample 1 results, the modular questionnaire was dropped for Samples 2a and 3 in favour of the singular questionnaire.

The propensity model worked well in classifying sample members into high-/low-response categories. However, the incentive treatment did not seem to work as well as expected. The model-directed treatment tended to slightly increase response rates for the low-propensity group and to slightly lower response for the high-propensity group. This latter effect might have been anticipated, given that the total incentive was higher for the high-propensity group in the standard incentive package. These results suggest that the total incentive for both the propensity groups might have been inadequate. Additional details on the experimental design can be found in the study by Biemer et al. (2020).

4.2 | Experiments in Samples 2a and 3

4.2.1 | Revised model-guided incentives

In preparation for fielding Samples 2a and 3, an additional analysis of the low-propensity group in Sample 1 was conducted. Our investigation concluded that this group comprises two somewhat disparate Wave IV subgroups: non-respondents and reluctant respondents. The Sample 1 response rate for the former group was only 17% compared with 35% for the latter group. This revelation led to a refinement of the model-directed protocol that split the low-propensity group further into two groups referred to as the *lower* low-propensity group (Wave IV non-respondents) and a *higher* low-propensity group (Wave IV reluctant respondents). In Samples 2a and 3, the former group was offered an incentive package with a monetary value of $100, whereas for the latter group, the monetary value was $65. The high-propensity group received a package valued at $55—the same as the package offered to the high-propensity group in Sample 1.

Another somewhat unexpected result was the small percentage of respondents who opted to respond by paper—less than 3% responded by paper compared with 36% in the Biemer et al. (2018) study that used a similar protocol. Thus, to save costs with little risk of higher non-response, the paper questionnaire including the additional $20 'push to web' payment was dropped in Samples 2a and 3. Although sample members could still request a paper questionnaire if they preferred that form of response, less than 1% of respondents did.

The incentive structure was as follows:

- High-Propensity (Low Incentive Package) Group
  - Total incentive = $55
    - $0–$10 prepaid incentive (see Section 4.2.2)
    - Promised the difference between $55 and the prepaid incentive for completing the questionnaire
- Higher Low-Propensity (Medium Incentive Package) Group
  - Total incentive = $65
    - $0–$10 prepaid incentive (see Section 4.2.2)
    - Promised the difference between $65 and the prepaid incentive for completing the questionnaire

- Lower Low-Propensity (High Incentive Package) Group
  - Total incentive = $100
    - $0–$10 prepaid incentive (see Section 4.2.2)
    - Promised the difference between $100 and the prepaid incentive for completing the questionnaire

At the completion of Phase 1 for Samples 2a and 3, the response rate for the lower low-propensity group was 28%, and for the higher low-propensity group, it was 44%, for a combined response rate of 36.7%. Comparing this with the low-propensity group under the singular questionnaire version in Table 2, this is an increase in response rate of 4.7 percentage points (36.7%–32.0%) for the model-directed incentive and 5.2 percentage points (36.7–31.5) for the standard incentive, both of which are highly significant. Thus, the higher incentives significantly increased response rates for Samples 2a and 3 compared with Sample 1. This increase is primarily a result of the significant increase in response rates for the lower low-propensity group, which was only 22% in Sample 1.

4.2.2 | Experiments with the prepaid incentive

In Sample 2a, an additional experiment was conducted to increase response rates. Arrayed in a 2x2 randomized factorial design, the first factor included a $10 prepaid incentive in the invitation mailing versus no prepaid incentive. The second factor sent a pre-notice letter in the form of a greeting card thanking the sample member for being part of Add Health for more than 20 years versus no pre-notice letter. The four treatments were assigned to four approximately equal random samples as follows: (a) greeting card pre-notice followed by invitation letter with $10 cash incentive, (b) $10 prepaid incentive only, (c) greeting card pre-notice only and (d) neither a pre-notice letter nor $10 prepaid incentive (control). Each treatment was embedded within propensity strata. Not investigated in this study was the effect of the greeting card visual format and message compared with a more traditional pre-notice postcard or plain letter.

Results of the 2x2 factorial design were examined over the course of Sample 2a data collection (a 12-month period). The impact of each treatment on response rates was evaluated at six time points in data collection: 1 week, 2 weeks, 1 month, 3 months, 6 months and 12 months. Logistic regression models were run at each time point to test main effects and two-factor interaction effects (Griggs et al., 2019). Unweighted response rates for each treatment are shown in Table 3.

Compared with those in the control group, sample members who received at least one of the treatments were significantly more likely to respond throughout data collection. This is consistent with the literature that suggests pre-notice letters and prepaid incentives are effective for increasing response rates (see, e.g. Dillman et al., 2014). In the regression, the interaction term was significant for only the first three time points (1 week, 2 weeks and 1 month). The models run at 3, 6 and 12 months show no significant interaction effect. In fact, there is no significant difference in response rates at 12 months among the three treatment groups. However, the rates of all three are significantly higher than the control group's rate.

Thus, it appears that early in data collection, sending a pre-notice greeting card and a pre-incentive significantly increased the preliminary response rates over sending only one of these items. However, except for the increase in early response rates, there seems to be little added benefit to the $10 incentive. Further, the cost of sending a greeting card is considerably lower than the cost of sending a $10 prepaid incentive. For these reasons, only the greeting card (condition c) was used in Sample 3.

TABLE 3 Response rate (%), by treatment at given time intervals during data collection.

Experimental Group	1 week	2 weeks	1 month	3 months	6 months	12 months
(a) Greeting card pre-notice and $10 prepaid incentive	20.2	27.2	36.4	43.6	51.7	52.3
(b) $10 prepaid incentive only	15.9	24.9	33.5	40.9	51.3	53.0
(c) Greeting card pre-notice only	14.3	23.1	31.9	39.9	49.2	51.0
(d) No pre-notice letter nor prepaid incentive (control)	6.8	14.6	22.5	31.4	43.4	45.6

4.3 | The benefits of the phase 2 NRFU

Following Phase 1 in mixed-mode samples, a random sample of non-respondents was followed up in person by field interviewers. This NRFU component of data collection is referred to as Phase 2. Non-respondents who were not selected for Phase 2 were automatically classified as *final* non-respondents. For cost efficiency, Phase 2 NRFU and Sample 2b field interviewing were concurrent using the same interviewing staff where possible. For Sample 1, the Phase 2 sampling rate was about 50%. Non-respondents were selected with probabilities inversely proportional to their Phase 1 sample base weights (see Biemer et al., 2020, for more details). This unequal probability sampling approach was successful at reducing the unequal weighting effects typically associated with two-phase sampling.

It can be shown (see, e.g. Groves & Heeringa, 2006) that this two-phase approach provides an effective response rate of $\rho = \rho_1 + (1 - \rho_1)\rho_2$, where ρ_1 and ρ_2 are the Phase 1 and Phase 2 response rates respectively. For example, in Sample 1, the response rate for Phase 1 was about 39%, and for Phase 2, it was about 32%. This equates to an 'effective' response rate of about 58%. The word 'effective' means that the two-phase data collection protocol produces essentially the same non-response bias risk (based on the non-response rate) as a single-phase data collection protocol with the same response rate, or $\rho = 58\%$ in this example. One disadvantage of following up only half of the non-respondents is that the number of respondents produced is also only about half compared with a 100% NRFU. However, an important advantage is substantially reduced cost as a result of the approximately 50% reduction of fieldwork.

Following an evaluation of the mean squared errors (MSEs) of the estimates for Sample 1 (discussed as follows), it was determined that the proportion of NRFU cases selected from Phase 1 non-respondents could be reduced to as low as 30% in Samples 2a and 3, thus providing considerable cost savings. However, doing so would also reduce the total number of completed interviews for the survey and the statistical power for subsequent data analysis. For that reason, the 50% sampling rate was retained for Samples 2 and 3.

One question considered in our analysis is whether the addition of Phase 2 was worth the cost. How much accuracy is gained by adding Phase 2 to the Phase 1 sample? To provide an answer, we compared the MSEs for two estimators—one based solely on the data from Phase 1 and the other based on the data from both phases. Denote by Y a population total of interest, by $\widehat{Y}_1$ the Phase 1 (only) estimator and by $\widehat{Y}_{1+2}$ the estimator using both phases. We assume that $\widehat{Y}_1$ is adjusted for non-response and coverage error using standard post-survey weighting adjustment procedures but that no Phase 2 data were used in the weighting adjustments. By contrast, $\widehat{Y}_{1+2}$ takes maximal advantage of the data from Phases 1 and 2 to reduce bias and variance. To compute and compare

the two MSEs, we assume that the non-response bias in $\hat{Y}_{1+2}$ is small relative to the corresponding non-response bias in $\hat{Y}_1$. Then, $\hat{Y}_1 - \hat{Y}_{1+2}$ can be used as an estimator of the approximate bias in $\hat{Y}_1$.

To compute $\text{MSE}(\hat{Y}_1)$, we apply the estimator in Potter et al., (1990), given by

$$\text{MSE}(\hat{Y}_1) = (\hat{Y}_1 - \hat{Y}_{1+2})^2 - \text{var}(\hat{Y}_{1+2}) + 2\hat{\rho}_{12}\sqrt{\text{var}(\hat{Y}_1)\text{var}(\hat{Y}_{1+2})}, \tag{1}$$

where $\hat{\rho}_{12}$ is an estimate of the correlation between $\hat{Y}_1$ and $\hat{Y}_{1+2}$, and var $(\cdot)$ denotes the estimator of the variance for the quantity within the parentheses. Like Potter et al. (1990), we conservatively assume $\hat{\rho}_{12} = 1$; however, we investigated the sensitivity of our results to this assumption, described as follows. Note that, by assumption, an estimator of $\text{MSE}(\hat{Y}_{1+2})$ is simply its estimated variance denoted by $\text{var}(\hat{Y}_{1+2})$.

We computed estimates of $\text{MSE}(\hat{Y}_1)$ and $\text{MSE}(\hat{Y}_{1+2})$ for Sample 1 data for 10-questionnaire variables thought to represent the most salient substantive variables in Add Health; variables such as *hours at work/week*, *number of live births*, *general health*, *alcohol use*, *sexual orientation* and *number of arrests*. The average of the MSE estimates over these variables, denoted by $\overline{\text{MSE}}_1$ and $\overline{\text{MSE}}_{1+2}$, respectively, provides an overall assessment of the MSEs for the items in the survey. If $\overline{\text{MSE}}_1 > \overline{\text{MSE}}_{1+2}$, one can conclude that the average accuracy of the estimates was improved by Phase 2; otherwise, Phase 2 provided no improvement on average.

The number of Phase 1 and Phase 2 respondents in Sample 1 is 3,665 and 1,110, respectively, for a combined total of 4,775 respondents. Across the 10 variables, $\overline{\text{MSE}}_1$ ranged from 8.0 to 8.6 depending on the value of $\hat{\rho}_{12}$ in Equation (1), which was allowed to range from 0.5 to 1.0. By comparison, $\overline{\text{MSE}}_{1+2}$ is only 1.3. To put this in relative terms, the mean value of the 10 estimates is 46, which implies that the average relative MSE for the Phase 1 estimates (i.e. $\overline{\text{MSE}}_1/46$) is 19%. The addition of Phase 2 reduces the average relative MSE from 19% to only 3%, which is a substantial improvement in accuracy.

Another question addressed in the Sample 1 analysis is how to balance Phase 2 costs and error in Samples 2a and 3. Although reducing the size of the Phase 2 sample would increase the variance of the adjusted estimates, it would save costs and still provide some reduction in total error. Thus, we computed $\overline{\text{MSE}}_{1+2}$ for a range of sample sizes from 10% to 50% of the Phase 1 non-respondents in Sample 1. The analysis suggested that reducing the Phase 2 sampling rate from 50% to 25% would approximately double $\overline{\text{MSE}}_{1+2}$. However, doubling the MSE would still result in a considerable reduction in total error compared with the $\overline{\text{MSE}}_1$. In fact, the ratio of $\overline{\text{MSE}}_{1+2}$ to $\overline{\text{MSE}}_1$ is still about 30% after reducing the sampling rate to 25%, which represents a 70% reduction in average MSE.

4.4 | Final response rates

Table 4 provides the response rates for Samples 1, 2a, 2b and 3 individually and combined. The formulas for the weighted response rates used in this table are

$$\rho_{\omega 1} = \frac{\sum_{i=1}^{n_1} \omega_i r_{1i}}{\sum_{i=1}^{n_1} \omega_i}, \ \rho_{\omega 2} = \frac{\sum_{i=1}^{n_2} \omega_i r_{2i}}{\sum_{i=1}^{n_2} \omega_i} \quad \text{and} \quad \rho_{\omega} = \frac{\sum_{i=1}^{n} \omega_i (r_{1i} + r_{2i}/\pi_{2i})}{\sum_{i=1}^{n} \omega_i}, \tag{2}$$

TABLE 4 Response rates, by phase: sample and overall

	Sample Size	Completed Interviews	Response Rate (Weighted)	Response Rate (Unweighted)
Sample 1				
Phase 1	7,867	3,665	42.4	46.6
Phase 2	2,430[a]	1,110	45.4	45.7
Combined	7,867	4,775	69.0	71.0[b]
Sample 2a				
Phase 1	2,683	1,497	54.4	55.8
Phase 2	667[c]	220	32.9	33.0
Combined	2,683	1,717	69.0	70.4[b]
Sample 3				
Phase 1	7,513	3,905	50.7	52.0
Phase 2	2,068[c]	801	38.2	38.7
Combined	7,513	4,706	69.2	70.5[b]
Sample 2b				
Sample 2b	1,526	1,102	71.7	72.2
All Samples				
Phase 1	19,589	10,169	49.5	51.9
Phase 2	5,165	2,131	40.8	41.3
Combined	19,589	12,300	69.3	71.8[b]

[a]NRFU sampling rate 50%

[b]Effective response rate computed as $\rho = \rho_1 + (1 - \rho_1)\rho_2$

[c]NRFU sampling rate is 55%.

where ω_i is the Add Health Wave I final weight for the ith sample unit, $\rho_{\omega 1}$, $\rho_{\omega 2}$ and ρ_{ω} are the Phase 1, Phase 2 and combined weighted response rates respectively; n_1, n_2 and n are the number of eligible sample members selected for Phases 1, 2 and for the total starting sample respectively; $r_{1i} = 1$ if the ith case responds in Phase 1 and is 0 otherwise; $r_{2i} = 1$ if the ith case responds in Phase 2 and is 0 otherwise; and π_{2i} is the following Phase 2 inclusion probability. The unweighted response rate has the same forms except ω_i and π_{2i} are replaced by 1.

As shown in the table, the weighted response rate for Sample 2b was around 72%. The weighted response rate for the mixed-mode samples (both phases combined) was about 69% for an overall weighted response rate of 69.3% (data not shown). Thus, we were able to achieve very nearly the same weighted response rate using mixed-mode data collection as in-person for substantial cost savings.

Note that the Phase 1 unweighted response rate for Sample 1 was 46.6%, which increased to 55.8% in Sample 2a and 52.0% in Sample 3 because of the improvements described in Sections 4.1 and 4.2. However, counteracting this improvement was a drop in Phase 2 response rates from 45.7% to 33.0% in Sample 2a and 38.7% in Sample 3. As a result, the improvement in the combined response rates across the three samples is less: from 60.7% in Sample 1 to 64.0% in Sample 2a to 62.6% in Sample 3.

After weighting, the combined response rates for the mixed-mode samples are all about 69%, which is <2 points off the in-person response rate. The unweighted (effective) response rate for all four samples combined is about 72%. By contrast, the response rate in Wave IV, which was all in-person, was about 80%.

5 | MODE-EFFECT ANALYSIS

This section provides a short summary of some results from a comprehensive analysis conducted for Wave V to assess the effect on measurement quality of transitioning Add Health from an in-person to a mixed-mode data collection design. Additional technical details and results of the methodology are reported in the study by Biemer et al. (2021).

As described in Section 4, the parallel sample (i.e. Sample 2b) implemented the same in-person interviewing methodology as was used in the prior Add Health waves. This design provides an opportunity to evaluate mode effects (i.e. the change in accuracy of Wave V estimates compared with that of prior waves) using Markov latent class analysis (MLCA). Sufficient conditions for MLCA model identifiability require a parallel sample plus a minimum of three Add Health waves, provided by Waves III, IV and V. This section describes the application of MLCA to evaluate and compare the accuracy levels of Wave III, IV and V cross-sectional and longitudinal estimates and presents the results for a small but representative subset of questionnaire items.

The mode analysis is divided into two parts. First, the so-called differential mode effect (DME) is estimated, defined as the difference between a mixed-mode sample (Sample MM) estimate and the corresponding in-person mode sample (Sample 2b) estimate. As will be shown, the DME has implications for the validity of cross-sectional and longitudinal analyses. The second part of the analysis decomposes DME into bias terms for non-response and measurement error, then combines these bias estimates with estimates of variance to obtain estimates of the total MSE. A description of the estimation methodology will be presented, followed by a presentation of some results, then a discussion of the implications of the results for the cross-section and longitudinal data analyses.

A simple indicator of the existence of a Wave V mode effect is the DME defined as $\text{DME} = \bar{y}_{MM} - \bar{y}_{2b}$, where $\bar{y}_{\text{MM}}$ is the mean of the combined mixed-mode sample (Sample MM), and $\bar{y}_{2b}$ is the mean of the in-person sample (Sample 2b). Because $\bar{y}_{\text{MM}}$ and $\bar{y}_{2b}$ are based on random samples of the same population, they should agree (apart from sampling error) when there are no mode effects. Thus, the magnitude of the DME can be used as an indicator of the magnitude of the existence of non-sampling errors that vary by mode. Our analysis focuses on three error sources that can contribute to the mode effect (viz., measurement error, non-response error and sampling error). By itself, the DME is insufficient for determining which sample produces more accurate estimates. That requires a comparison of the mode-specific error components, which is the aim of the MLCA described in the next section.

5.1 | Methodology

MLCA is a model-based approach for estimating the measurement error parameters in longitudinal, categorical data. As described in the study by Biemer (2011), the measurement error in a sample proportion is a function of the misclassification probabilities associated with the measurement process (i.e. the false negative and false positive error rates in the case of dichotomous data). The probability of not observing a false positive response is referred to as *specificity*,

whereas the probability of not observing a false negative response is referred to as *sensitivity*. Because of the Wave V parallel sample design, MLCA can be used to estimate specificity and sensitivity directly from Add Health data without requiring 'gold standard' measurements or external estimates assumed to be error free.

Because it uses data from Waves III, IV and V, MLCA can be applied only to items that were collected in all three of these waves. Moreover, the item constructs across the three waves should be identical (i.e. no substantive wording changes in the item questions). A total of 22 items were selected that span the range of sensitive, factual and subjective topics. Because of space limitations, only the 10 items shown in Table 5 will be presented in this section. These items convey the breadth of results from the full analysis. Items excluded from Table 5 include religious affiliation, arrests, physical activity, political ideology, volunteerism and visual impairment (see Biemer et al., 2021, for those results).

The first column in Table 5 provides the questionnaire wording for each item. In all prior waves and in Wave V, the five italicized items were collected by CASI in the in-person mode to reduce the risks of social desirability effects (see, e.g. Tourangeau et al., 2000). Items not italicized were obtained by interviewers. Column 2 shows the original number of categories for each of the selected items, and Column 3 lists the abbreviation we will use to refer to the item. As shown in Column 4, all the items are dichotomous or were recoded to be dichotomous.

A path diagram for the basic model implemented in the MLCA is shown in Figure 2. In this figure, S is a dichotomous variable denoting the sample where $S = 1$ for Sample MM and $S = 2$ for Sample 2b. For $t =$ Wave III, IV and V, M_t denotes the mode of data collection used (1 for web/paper, 2 for in-person), X_t (= 0 or 1) denotes the (dichotomous) latent construct and Y_t (= 0 or 1) denotes the corresponding (dichotomized) survey indicator of X_t. The term $S \times M_5$ in the first box

TABLE 5 Add Health questionnaire items used in the analysis

Item Wording	Original No. of Categories	Abbreviation	Recoded Item
In general, how is your health?	5	Goodhlth	In good health? 1(2) = yes(no)
Which of the following best describes your current health insurance situation?	14	Insure	Has insurance? 1(2) = yes(no)
In the past 12 months, have you had a dental examination by a dentist or dental hygienist?	2	Dental	1(2) = yes(no)
During the past 7 days, I felt that I could not shake off the blues, even with help from my family and friends.	4	Blue	Ever felt blue? 1(2) = yes(no)
During the past 7 days, I felt sad.	4	Sad	Ever felt sad? 1(2) = yes(no)
Have you ever had vaginal intercourse?	2	Intercourse	1(2) = yes(no)
Are you romantically attracted to females?	2	Att_Fem	1(2) = yes(no)
Are you romantically attracted to males?	2	Att_Mal	1(2) = yes(no)
Have you ever smoked cigarettes regularly—that is, at least one cigarette every day for 30 days?	2	CigUse	1(2) = yes(no)
During the past 12 months, have you ever seriously thought about committing suicide?	2	Suicide	1(2) = yes(no)

Note: Items in *italics* were collected by CASI in in-person mode.

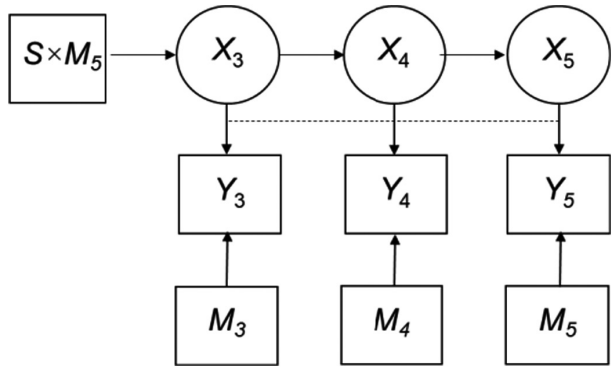

FIGURE 2 Path diagram for the basic MLCA model used for the analysis

in Figure 2 represents the effect of partitioning the full sample into four subsamples defined by S and M_5. The model essentially specifies that the prevalence of X_3 is allowed to vary across four subpopulations that received mixed and in-person modes at Wave V.

To convert Figure 2 into a mathematical model requires additional notation. Following Biemer (2011), probabilities are represented by π, where superscripts denote random variables and subscripts denote their values. For example, $Pr(X_t = 1)$, $Pr(Y_t = 1 \mid X_t = 0)$ and $Pr(Y_t = 1 \mid G = g, X_t = 0)$ are denoted by $\pi_1^{X_t}$, $\pi_{1|0}^{Y_t|X_t}$ and $\pi_{1|g0}^{Y_t|GX_t}$ respectively. A hat over a parameter denotes its estimator; for example, the prevalence of the latent variable at Wave III, $\pi_1^{X_3}$, is estimated by $\hat{\pi}_1^{X_3}$. The likelihood kernel for the model in Figure 2 is given by the following expression for the joint probability of an observation in an arbitrary cell of the joint classification table $S \times M_3 \times M_4 \times M_5 \times Y_3 \times Y_4 \times Y_5$:

$$\pi_{sm_3m_4m_5y_3y_4y_5}^{SM_3M_4M_5Y_3Y_4Y_5} = \pi_{sm_3m_4m_5}^{SM_3M_4M_5} \sum_{x_3x_4x_5} \left(\pi_{x_3|sm_5}^{X_3|SM_5} \pi_{x_4|x_3}^{X_4|X_3} \pi_{x_5|x_4}^{X_5|X_4} \right) \left(\pi_{x_3|m_3x_3}^{Y_3|M_3X_3} \pi_{x_4|m_4x_4}^{Y_4|M_4X_4} \pi_{x_5|m_5x_5}^{Y_5|M_5X_5} \right), \quad (3)$$

where $\Sigma_{x_3x_4x_5}$ denotes the summation over the latent variables in the model.

Several constraints are imposed on this model to achieve identifiability and to specify the structural 0s in the observed table. The first identifiability constraint (represented by the dotted line in Figure 2) sets $\pi_{y_3|m_3x_3}^{Y_3|M_3X_3} = \pi_{y_4|m_4x_4}^{Y_4|M_4X_4} = \pi_{y_5|m_5x_5}^{Y_5|M_5X_5}$, which states that misclassification probabilities for Waves III, IV and V are the same for observations collected by the in-person mode. This restriction seems plausible because the in-person interview protocols, questionnaire wordings and respondents are the same across waves; thus, risks of misclassification for these respondents should at least be similar. The Markov assumption is imposed by setting $\pi_{x_5|x_4x_3}^{X_5|X_4X_3} = \pi_{x_5|x_4}^{X_5|X_4}$. This assumption, which is also required for identifiability, states that Wave IV to Wave V transitions (i.e. a true Wave IV positive becomes a true Wave V negative and vice versa) are not influenced by an individual's Wave III state. The final assumption required for identifiability is inter-wave local independence represented in Figure 2 by the absence of an arrow between Y_t for $t = 3,4$ and 5. It is represented in Equation (3) by expression for the joint probability of Y_3, Y_4, Y_5, given X_3, X_4, X_5 in the last parenthetical term of the equality, ignoring M_3, M_4 and M_5.

Two types of structural 0s are imposed on the model. First, recall that the mode at Wave 5 was in-person ($M_5 = 2$) for Samples MM (Phase 2) and 2b, and it was web/paper ($M_5 = 1$) for Sample MM (Phase 1). Because the mixed mode was not used in Waves III and IV, M_3 and M_4 cannot be 1 and structural 0s must be imposed in the model for these cells. Second, note that the combinations $(S, M_5) = (1,1)$ and $(S, M_5) = (1,2)$ correspond to Phases 1 and 2, respectively, of Sample MM—the

first conducted by web/paper and the second by in-person interviewing. Further,$(S, M_5) = (2,2)$ corresponds to Sample 2b, which was entirely in-person. Because the web/paper mode was not used for Sample 2b, structural 0s must also be imposed for $(S, M_5) = (2,1)$.

Finally, another feature incorporated in the MLCA is a non-response component representing the random selection of Phase 1 non-respondents for Phase 2 NRFU denoted by $\pi^{R|SX_5}_{r|sx_5}$, where $R = 1$ for all respondents in the S, and $R = 0$ for Phase 1 non-respondents who were not selected for Phase 2. Other Phase 1 non-respondents were excluded from the analysis. Retaining these Wave V, Sample MM, Phase 1 non-respondents in the analysis improves the precision of the estimates because otherwise the Wave III and IV data for these non-respondents would be excluded in the analysis. Additional details of this modelling approach can be found in the study by Biemer et al. (2021).

5.2 | Results of the mode-effects analysis

As described in the study by Biemer et al. (2020), survey weights were developed separately for Samples MM, 2b and the combined sample that included adjustments for non-response and coverage. Using an approach like that of Singh et al. (2003), the weights were calibrated to sampling frame and questionnaire variables and reflected the two-phase data collection approach used for Sample MM. The weights for Sample 2b mirrored the weighting approach that was used for Waves III and IV. Thus, the non-sampling error properties of the Sample 2b estimates should be very similar to the corresponding estimates from Waves III and IV. In addition, comparisons between the fully weighted Sample MM and 2b estimates will be used to evaluate the potential bias in comparisons of Wave V and prior wave estimates.

The second column of Table 6 provides the relative DMEs defined as RDME = DME/$\bar{y}_V$, where $\bar{y}_V$ is the full Wave V sample. The symbols 'a', 'b' and 'c' in the table indicate significance ($\alpha = 10\%$) for three hypotheses: (a) |RDME| = 0, (b) |RDME| ≤ 0.05 and (c) |RDME| ≤ 0.10. The absolute RDME significantly exceeds 10% for two items (Blue and Sad) and significantly exceeds 0 for one item (Intercourse).

For the full table of 22 items, the average DME was only 0.41 percentage points; however, there was considerable variability across items, from −12.3 percentage points to 11.6 points. The average absolute DME was about 3.5 points, and the average absolute RDME was 9.6%. In general, the mixed-mode sample produced estimates that were about 2.6% higher than the in-person sample.

Table 6 shows the sensitivities and specificities for the two modes: web/paper (Columns 2 and 4 respectively) and in-person (Columns 3 and 5 respectively), estimated using the Latent GOLD® 5.1 software (Vermunt & Magidson, 2015). The last two columns show the ratios of web/paper to in-person sensitivities and specificities respectively. In Table 6, except for three items (which are labelled '*ns*' in Column 1), almost all items significantly differ by mode on sensitivity, specificity or both (α = 10%). For the full set of 22 estimates, the sensitivity ratio ranged from 0.88 to 1.94 with an average of 1.05. Thus, the web/paper mode is slightly more sensitive, on average.

For specificity, the ratios ranged from 0.84 to 1.06 with an average of 0.98, indicating slightly less specificity for web/CAPI. Finally, the last two columns of Table 6 provide the measurement accuracy rates (Acc) for web/paper and in-person modes, respectively, where Acc is defined as the probability the survey response agrees with an individual's true classification. In the full study, the 22 items were equally split on mode accuracy, with 11 items having higher web/paper accuracy and 11 items having higher in-person accuracy. However, the average level of accuracy (for all 22 items in the study) was slightly higher for the web/paper mode: 88.51% versus 87.76%.

TABLE 6 Comparing the estimates of item sensitivity, specificity and accuracy for web/paper (MW) and in-person (IP) modes (All Entries Are In Percentages)

Item	RDME	Web/Paper Sensitivity (Sen_{MW})	In-Person Sensitivity (Sen_{IP})	Web/Paper Specificity (Spe_{MW})	In-Person Specificity (Spe_{IP})	Acc_{MW}	Acc_{IP}
Goodhlth	−0.66	99.3	95.9	53.9	53.2	88.45	85.69
Insure	−0.08	98.6	97.6	53.9	57.3	92.29	91.91
Dental	−2.23	83.0	83.2	70.0	70.3	78.11	78.34
Blue	29.45[b]	70.1	49.6	87.3	91.1	82.81	80.26
Sad	23.66[b]	84.9	64.8	73.5	80.7	78.34	73.95
Intercourse	−2.13[a]	98.6	98.8	82.2	97.9	98.00	98.77
Att_Fem	−1.14	97.2	98.5	97.1	96.2	97.15	97.41
Att_Mal	0.67	98.7	99.1	99.0	98.7	98.85	98.90
CigUse	−7.88	93.6	95.3	98.9	94.4	96.33	94.84
Suicide	11.51	65.6	33.7	97.2	96.7	95.36	93.03

Note: Items in *italics* were collected by CASI in in-person mode.

[a]|RDME| = 0 rejected at 10% significance level.

[b]|RDME| ≤ 0.10 rejected at 10% significance level.

We also computed the biases of the weighted mean estimates for Samples MM, 2b and the full Wave V sample (referred to as Sample V) and their standard errors. In each case, the total bias was estimated from the MLCA as the sum of measurement error bias and non-response bias. For example, measurement bias for sample s is estimated by $\hat{B}_{ME,s} = (\bar{y}_s - \bar{x}_s)$, where $\bar{y}_s$ is the weighted sample mean of the observations and $\bar{x}_s$ is the weighted sample mean of the latent construct, X, estimated from the MLCA. Non-response bias is estimated by $\hat{B}_{NR,s} = (\bar{x}_s - \bar{x}_V)$, where it is assumed that $\hat{B}_{NR,V} = 0$; that is, the weighted mean of X (which, under the MLC model, is measurement error free) over the entire sample, S_V, is unbiased (or minimally biased). This assumption is plausible in the sense that $\bar{x}_V$ is the best estimator available given its size and the extensive non-response adjustments that were applied to it. To the extent this assumption is violated, the magnitude of the non-response bias estimates will likely be understated.

The following summarizes the key results from the mode analysis:

- Regarding |RDME|, eight items are significantly larger than 0; five of these significantly exceed 5%, and four significantly exceed 10%.
- On average, the bias contribution from non-response is quite small for Sample MM (−0.08%) and larger for Sample 2b (2.42%). However, this finding may be an artefact of the method for estimating non-response bias, which tends to favour Sample MM.
- The magnitudes of total bias ($\hat{B}_{ME,s} + \hat{B}_{NR,s}$) are much higher for Sample 2b than for Sample MM. Two items had total biases significantly exceeding 10% in Sample MM compared with six items for Sample 2b.
- For estimates based on the full sample, six items have total biases significantly exceeding 10%—the same six items that were noted in the previous bullet for Sample 2b.

For 16 of the 22 items in the analysis, the Wave V design produced smaller relative root mean squared errors (RRMSEs) than the traditional design it replaced. If the non-response

bias component (which, as previously noted, tends to favour Sample MM) is removed from the calculations, the Wave V design's RRMSEs are smaller in only nine out of 22 items. Still, these results are surprising given in-person interviewing has long been regarded as the ideal mode for Add Health. It is also unexpected given that the response rates in prior waves were much higher than for Wave V; for example, the Sample 2b response rate, which is the best indicator of what a traditional Add Health design would produce, is 9 percentage points higher than the overall Wave V response rate. For weighted response rates, which are better indicators of the non-response bias risk in weighted estimates, the gap closes to only 2.7 percentage points—69.3% for the full Wave V sample versus 72% for Sample 2b—suggesting very similar non-response bias risks. We might also expect the more elaborate post-survey weighting methodology applied to the full Wave V sample to be more effective at reducing non-response bias than traditional Add Health weighting approaches. Finally, on average, the RRMSEs for the traditional design are about 60% higher than for Wave V design, even after excluding the non-response bias component.

6 | LESSONS LEARNED AND RECOMMENDATIONS

Funding reductions combined with increasing data collection costs motivated the transition of Add Health at Wave V (2016–18) from traditional in-person interviewing to a more cost-effective approach where about 70% of the observations were collected via the web. A responsive design was implemented that featured experimentation in the early samples to inform methodological improvements in latter samples. The full sample was partitioned into four nationally representative subsamples. Samples 1, 2a and 3 used a two-phase design with web/paper in Phase 1 and in-person NRFU in Phase 2. Overall, about half of the Phase 1 non-respondents were selected in Phase 2. Sample 2b was conducted in person and served as a sort of control group for evaluating mode effects. Experiments performed during Sample 1 determined the questionnaire design and the incentive structures that varied according to estimated response propensities in the mixed-mode samples. Additional experiments in Sample 2a led to improvements in the introductory mailings in Sample 3.

6.1 | Lessons learned

Perhaps the most important lesson learned in this transition wave is that is possible to successfully transition a panel survey with a long history of traditional in-person interviewing to the web/paper mode with no apparent deterioration in data quality. In fact, by all accounts, Wave V data quality appears to have generally improved compared with that of prior waves for cross-sectional data analysis.

However, this cannot be said for longitudinal analyses, for which large mode effects risk confounding estimates of prior wave to Wave V change. Mode effects when transitioning from an interview-assisted to a self-administered mode may be unavoidable in surveys that deal with sensitive phenomena. The findings of large mode effects in Add Health accentuates the importance of incorporating some provision for evaluating and compensating mode effects in any panel survey contemplating such a redesign.

The following are a few additional lessons learned that may have implications for panel surveys in general:

1. *The model-guided incentive structure can substantially increase response rates, provided that the incentive payouts are thoughtfully and appropriately designed.*

The propensity model based on prior wave response patterns successfully identified the low-responding sample members, but the $75 incentive employed in Sample 1 appeared to be inadequate for substantially increasing their response propensities. The strategy employed in Samples 2a and 3 increased response rates by almost 5 percentage points for this group. Essentially, the low-response propensity group was further partitioned into two subgroups: one lower propensity and one higher propensity. The former was offered $100, whereas the latter was offered $65 in total incentives.

2. *Dividing the questionnaire into two modules that are fielded sequentially was not effective in increasing response rates. In fact, response rates were slightly lower for the modular questionnaire design.*

This finding suggests that dividing the questionnaire into two modules may pose a greater burden to respondents than completing a single questionnaire of an equivalent length. In addition, a web/paper questionnaire of about 50 min may not be an important factor for panel survey members who, in four prior waves, completed a 90-min questionnaire.

3. *The Choice + protocol, which offers two options—(a) respond by paper questionnaire or (b) for an additional $20 promised incentive, respond online—did not appear to be an important factor in pushing respondents to the web.*

In Sample 1, response by paper questionnaire was only 3%, substantially lower than the 36% paper response reported for the Choice + protocol in the study by Biemer et al. (2018). This result suggests that very few Wave V respondents, most of whom are in their 30s with at least a high school education, are without internet access. Dropping the paper questionnaire mailings and the $20 'push to web' incentive saved substantial costs with no apparent reduction in response rates.

4. *The Phase 2 in-person NRFU, while adding substantial costs, appears to have been successful at significantly reducing non-response bias and improving the accuracy of the web/mail estimates.*

The average relative MSE for a representative selection of 10-questionnaire items was reduced from 19% to only 3% by the addition of the Phase 2 in-person NRFU sample. Apparently, the in-person mode was successful in obtaining responses from sample members with characteristics significantly different from those who responded to the web/mail survey.

5. *Other than increasing early response rates, there was no added benefit of sending a greeting card pre-notice and including $10 in the invitation letter. However, doing either of these performed significantly better than doing neither.*

This finding is consistent with the literature that suggests that pre-notice letters and prepaid incentives are both effective for increasing response rates (see, e.g. Dillman et al., 2014). Still, the absence of an additive effect of a prepaid incentive is unexpected given the compelling evidence

suggesting that prepaid incentives boost response rates (see, e.g. the meta-analysis by Göritz, 2006). Perhaps the null effect could be attributed, in part, to Add Health respondents, after 26 years in the program, trusting that the promised incentives will be delivered.

Regarding mode effects, the analysis found that data quality did not suffer after transitioning from in-person to the web/mail mode. These findings are good news but also largely unexpected based on the preponderance of literature suggesting in-person interviewing is the de facto gold standard data collection mode.

However, the bad news is the high risk of mode effects when comparing Wave V and prior wave estimates. Biemer et al. (2021) found that significant differences can occur in longitudinal change estimates about 60% of the time (13 of 22 estimates), purely as an artefact of the redesign. However, the true risks could be considerably higher because the benchmark sample in the mode analysis was small relative to a full wave of data as would be used in a full-sample longitudinal analysis. The scope of mode analysis is somewhat limited in that only 20 of hundreds of possible survey items in the Add Health instrument were examined, and the selected items may not be representative of all items. Nevertheless, these results raise questions regarding how data users should longitudinally analyse the Wave V data and interpret significant findings. Biemer et al. (2021) provide some guidance to analysts in that regard.

We believe the sequentially fielded sample replicates, the Phase 2 NRFU design and the in-person benchmark sample that comprise the Wave V design achieved their objectives in terms of cost efficiency, data quality and timeliness of reporting. Although longitudinal analysis will be challenging in the presence of the large mode effects, the benchmark sample will provide analysts with several options for compensating for mode effects, estimation and interpretation of the Wave V data (Biemer et al., 2021). Thus, these findings have important implications for other in-person, longitudinal surveys transitioning to web/paper.

6.2 | Recommendations

After discussing these findings and lessons learned, we offer five recommendations to other survey practitioners and methodologists who are considering transitioning an in-person panel survey to a mostly web/mail mode.

1. Use a two-phase design for the web/mail mode, where Phase 2 is an in-person follow-up of web/mail non-respondents. A minimum NRFU sample size of approximately 300–400 NRFU respondents seems sufficient for stable non-response adjustment factors based on the Add Health experience.
2. Include a parallel benchmark sample featuring in-person interviewing (like Wave V Sample 2b), and encourage data users to incorporate it into their analyses to assess the effects of mode changes.
3. Use a model-guided incentive structure that provides high-value incentives for the lowest propensity group.
4. Use a singular questionnaire for web/mail and in-person, and limit the length of the interview to about 1 h if possible.
5. Forgo including paper questionnaires in the web invitation mailings. Instead, encourage sample members to request a paper questionnaire if that is their preferred mode.

A limiting factor in the broad applicability of our research is the target population, which is a subpopulation of adults aged 34–42 who participated in up to five waves of a longitudinal survey. We caution that the aforementioned recommendations may not extend to the general population or other age groups. Nevertheless, we believe the results clearly demonstrate the advantages of the multi-sample, multi-phase benchmarked responsive design for informing the mode transition process. Recommendations 1 and 2 are especially important for other mode panel survey redesigns because the former substantially reduced non-response bias, and the latter provided valuable information on mode effects and opportunities for compensating for them in the data analyses.

ORCID

Paul P. Biemer https://orcid.org/0000-0003-2214-2707
Brian J. Burke https://orcid.org/0000-0001-5548-0388

REFERENCES

Abowd, J.M. & Zellner, A. (1985) Estimating gross labor force flows. *The Journal of Business & Economic Statistics*, 3, 254–283.

Aragon-Logan, E., Dean, E., Granger, B., Hinsdale, M., Hottinger, C. & Meehan, A. et al. (2010) Add Health Wave IV final report. RTI International. (Internal report).

Bianchi, A., Biffignandi, S. & Lynn, P. (2017) Web-face-to-face mixed-mode design in a longitudinal survey: Effects on participation rates, sample composition, and costs. *Journal of Official Statistics*, 33(2), 385–408.

Biemer, P.P. (2011) *Latent class analysis of survey error*. Hoboken, NJ: Wiley.

Biemer, P.P., Burke, B., Carson, C., Considine, K., Murphy, J. & Powell, R. (2020) Add Health Wave V final report. RTI International. (Internal report).

Biemer, P., Harris, K., Burke, B., Liao, D. & Halpern, C. (2021) Modeling mode effects for a panel survey in transition. In: Cernat, A. & Sakshaug, J. (Eds.) *Measurement errors in longitudinal surveys*. New York, NY: Oxford University Press.

Biemer, P.P. & Lyberg, L. (2003) Errors due to interviewers and interviewing. In: *Introduction to survey quality*. ch. 5, New York, NY: Wiley, pp. 149–187.

Biemer, P.P., Murphy, J., Zimmer, S., Berry, C., Deng, G. & Lewis, K. (2018) Using bonus monetary incentives to encourage web response in mixed-mode household surveys. *Journal of Survey Statistics and Methodology*, 6(2), 240–261.

Brown, J.L., Swartzendruber, A. & DiClemente, R.J. (2013) Application of audio computer-assisted self-interviews to collect self-reported health data: An overview. *Caries Research*, 47(Suppl. 1), 40–45.

Cernat, A. (2015a) Impact of mixed modes on measurement errors and estimates of change in panel data. *Survey Research Methods*, 9(2), 83–99.

Cernat, A. (2015b) The impact of mixing modes on reliability in longitudinal studies. *Sociological Methods & Research*, 44(3), 427–457.

Cernat, A., Couper, M.P. & Ofstedal, M.B. (2016) Estimation of mode effects in the Health and Retirement Study using measurement models. *Journal of Survey Statistics and Methodology*, 4(4), 501–524.

Crawford, S.D., Couper, M.P. & Lamias, M.J. (2001) Web surveys: Perceptions of burden. *Social Science Computer Review*, 19, 146–162.

de Leeuw, E.D. (2018) Mixed-mode: Past, present, and future. *Survey Research Methods*, 12(2), 75–89.

de Leeuw, E. & Collins, M. (1997) Data collection methods and survey quality: An overview. In: Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N. & Trewin, D. (Eds.) *Survey measurement and process quality*. New York: John Wiley.

de Leeuw, E.D. & de Heer, W. (2002) Trends in household survey non-response: A longitudinal and international perspective. In: Groves, R.M., Dillman, D., Eltinge, J. & Little, R. (eds) *Survey Non-Response*, New York, NY: Wiley.

de Leeuw, E.D. & van der Zouwen, J. (1988) Data quality in telephone and face-to-face surveys: A comparative meta-analysis. In: Groves, R., Biemer, P., Lyberg, L., Massey, J., Nichols II W. & Waksberg. J. (eds). *Telephone survey methodology*, New York, NY: Wiley.

Dillman, D.A. (2007) *Mail and internet surveys: The tailored design method*, 2nd edition. New York, NY: Wiley.
Dillman, D.A. (2009) Some consequences of survey mode changes in longitudinal surveys. In: Lynn, P. (Ed.) *Methodology of longitudinal surveys*. West Sussex, UK: Wiley, pp. 127–140.
Dillman, D.A. (2017) The promise and challenge of pushing respondents to the web in mixed-mode surveys. *Survey Methodology*, 12–001-X(43), 1. Available from http://www.statcan.gc.ca/pub/12-001-x/2017001/article/14836-eng.htm.
Dillman, D.A., Smyth, J.D. & Christian, L.M. (2009) *Internet, mail, and mixed-mode surveys: The tailored design method*, 3rd edition. Hoboken, NJ: Wiley.
Dillman, D.A., Smyth, J.D. & Christian, L.M. (2014) *Internet, phone, mail, and mixed-mode surveys: The tailored design methods*, 4th edition. Hoboken, NJ: Wiley.
Francis, J. & Laflamme, G. (2015) Evaluating web data collection in the Canadian Labour Force Survey. Paper presented at the Federal Committee on Statistical Methodology Research Conference, Washington, DC. Available from https://nces.ed.gov/fcsm/pdf/H2_Francis_2015FCSM.pdf.
Freedman, V.A. (2017) *The panel study of income dynamics' Wellbeing and daily life supplement (PSID-WB) user guide: Final release 1*. Ann Arbor, MI: Institute for Social Research, University of Michigan.
Galesic, M. & Bosnjak, M. (2009) Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73, 349–360.
Göritz, A.S. (2006) Incentives in web studies: Methodological issues and a review. *International Journal of Internet Science*, 1, 58–70.
Griggs, A.K., Powell, R.J., Keeney, J., Waggy, M., Harris, K.M., Halpern, C.T. et al. (2019) Research note: A prenotice greeting card's impact on response rates and response time. *Longitudinal and Life Course Studies*, 10(4), 421–432.
Groves, R. & Couper, M. (1998) *Nonresponse in household interview surveys*. New York, NY: Wiley.
Groves, R. & Heeringa, S.G. (2006) Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A*, 169(3), 439–457.
Gummer, T. & Roßmann, R. (2015) Explaining interview duration in web surveys: A multilevel approach. *Social Science Computer Review*, 33(2), 217–234.
Harris, K.M., Halpern, C.T., Whitsel, E.A., Hussey, J.M., Killeya-Jones, L., Tabor, J. et al. (2019) Cohort profile: The national longitudinal study of adolescent to adult health (Add Health). *International Journal of Epidemiology*, 48(5), 1415–1415k.
Health and Retirement Study (2019) *Data collection path diagram*. Ann Arbor, MI: Institute for Social Research, University of Michigan. Available from https://hrs.isr.umich.edu/data-products/collection-path.
Heerwegh, D. (2009) Mode differences between face-to-face and web surveys: An experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research*, 21, 111–121.
Heerwegh, D., Abts, K. & Loosveldt, G. (2007) Minimizing survey refusal and noncontact rates: Do our efforts pay off? *Survey Research Methods*, 1, 3–10.
Institute for Social and Economic Research (2020) *Understanding Society: Waves 1-10, 2009-2019 and Harmonised BHPS: Waves 1-18, 1991-2009, User Guide, 29 October 2020*. Colchester: University of Essex.
Loosveldt, G. & Beullens, K. (2013) 'How long will it take?' An analysis of interview length in the fifth round of the European social survey. *Survey Research Methods*, 7, 69–78.
Lynn, P. (2009) Sample design for Understanding Society. Understanding Society Working Paper Series 2009–01. Institute for Social and Economic Research.
McGonagle, K.A., Schoeni, R.F., Sastry, N. & Freedman, V.A. (2012) The panel study of income dynamics: Overview, recent innovations, and potential for life course research. *Longitudinal Life Course Studies*, 3(2), 268–284.
Murphy, J., Biemer, P. & Berry, J. (2018) Transitioning from face-to-face interviewing to a self-administration using adaptive, responsive, and tailored (ART) design principles. *Journal of Official Statistics*, 34(3), 625–648.
O'Reilly, J., Hubbard, M.L., Lessler, J.T., Biemer, P.P. & Turner, C. (1994) Audio and video computer-assisted self-interviewing: Preliminary tests of new technologies for data collection. *Journal of Official Statistics*, 10(2), 197–214.
Olson, K., Smyth, J. D., Horwitz, R., Keeter, S., Lesser, V., Marken, S. et al. (2021) Transitions from Telephone Surveys to Self-Administered and Mixed-Mode Surveys: AAPOR task force report. *Journal of Survey Statistics and Methodology*, 9(3), 381–411. https://doi.org/10.1093/jssam/smz062

Pickery, J. & Loosveldt, G. (2000) Modeling interviewer effects in panel surveys: An application. *Survey Methodology*, 26(2), 189–198.

Potter, F. (1990) A study of procedures to identify and trim extreme sampling weights. In Proc. American Statistical Association, Survey Research Methods Section, pp. 225–230.

Powell, R.J., Biemer, P.P., Cook, S.L., Considine, K.A., Halpern, C.T., Harris, K.M. et al. (2018) Two short or one long: An experiment comparing survey length vs. quantity of surveys. Paper presented at the 2018 Joint Statistical Meetings. Vancouver, Canada.

Sakshaug, J., Yan, T. & Tourangeau, R. (2010) Nonresponse error, measurement error and mode of data collection: Tradeoffs in a multi-mode survey of sensitive and non-sensitive items. *Public Opinion Quarterly*, 74(5), 907–933.

Singh, A.C., Iannacchione, V.G. & Dever, J.A. (2003) Efficient estimation for surveys with nonresponse follow-up using dual-frame calibration. Proc. American Statistical Association, Survey Research Methods Section, pp. 3919–3930.

Starick, R. & Steel, J. (2012) Does increased effort lead to a less representative response? Selected case studies from the Australian Bureau of Statistics. Paper presented at the European Conference on Quality in Official Statistics, Athens, Greece.

Tourangeau, R., Rips, L. & Rasinski, K. (2000) *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

University of Essex, Institute for Social and Economic Research (2019) *Understanding society: Innovation panel, waves 1-11, 2008-2018*, 9th edition. Colchester, UK: University of Essex.

Vermunt, J.K. & Magidson, J. (2015) *LG-Syntax™ user's guide: manual for latent GOLD® 5.0 syntax module*. Belmont, MA: Statistical Innovations Available from https://www.statisticalinnovations.com/wp-content/uploads/LGSyntaxusersguide.pdf.

Villar, A. & Fitzgerald, R. (2017) Using mixed modes in survey research: Evidence from six experiments in the ESS. In: Breen, M. (Ed.) *Values and identities in Europe: Evidence from the European social survey*. ch. 16, England, UK: Routledge, pp. 273–310.

Watson, N. & Wooden, M. (2009) Identifying factors affecting longitudinal survey response. In: Lynn, P. (Ed.) *Methodology of longitudinal surveys*. Hoboken, NJ: Wiley, pp. 151–181.

How to cite this article: Biemer, P.P., Harris, K.M., Burke, B.J., Liao, D. & Halpern, C.T. (2022) Transitioning a panel survey from in-person to predominantly web data collection: Results and lessons learned. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185, 798–821. https://doi.org/10.1111/rssa.12750