AN EXAMINATION OF AN INTERVIEWER-RESPONDENT MATCHING PROTOCOL IN A LONGITUDINAL CATI STUDY

BRADY T. WEST*

MICHAEL R. ELLIOTT

ZEINA MNEIMNEH

JAMES WAGNER

ANDY PEYTCHEV

MARK TRAPPMANN

This article presents results from an experimental study in Germany designed to test the effectiveness of a novel protocol for matching participants in a national panel survey with interviewers employing computer-assisted telephone interviewing (CATI) on selected sociodemographic features, including sex, age, and education. We specifically focus on the ability of the protocol to engender close matches between respondents and interviewers in terms of these features, using both theory and empirical evidence to suggest that this type of matching will improve

BRADY T. WEST is Research Associate Professor in the Survey Methodology Program (SMP), Survey Research Center (SRC), Institute for Social Research (ISR), University of Michigan, 426 Thompson Street, Ann Arbor, MI 48106, USA. MICHAEL R. ELLIOTT is Professor of Biostatistics at the School of Public Health and Research Professor in the Survey Methodology Program (SMP), Survey Research Center (SRC), Institute for Social Research (ISR), University of Michigan. ZEINA MNEIMNEH is an Assistant Research Scientist in the Survey Methodology Program (SMP), Survey Research Center (SRC), Institute for Social Research (ISR), University of Michigan. JAMES WAGNER is Research Associate Professor in the Survey Methodology Program (SMP), Survey Research Center (SRC), Institute for Social Research (ISR), University of Michigan. ANDY PEYTCHEV is a Senior Survey Methodologist in the Center for Survey Methodology, Survey Research Division, RTI International, 3040 E. Cornwallis Rd., Research Triangle Park, NC 27709, USA. MARK TRAPPMANN is Head of Research on the Labor Market and Social Security Study at the Institute for Employment Research, Regensburger Str. 104, 90478 Nuernberg, Germany and Professor of Survey Methodology at the University of Bamberg, Feldkirchenstr. 21, 96052 Bamberg, Germany. The authors thank the editors and two reviewers who improved the quality of the manuscript.

Financial support for the work of West, Elliott, and Wagner was provided by a grant from the National Institutes of Health (#1 R01 AG058599-01).

*Address correspondence to Brady T. West, Survey Methodology Program, 4118 Institute for Social Research, 426 Thompson Street, Ann Arbor, MI 48106, USA; E-mail: bwest@umich.edu.

cooperation rates in surveys employing CATI. We also focus on indicators of “success” at first contact (defined as a successful interview or establishment of an appointment for an interview) as a function of whether the matching protocol was in use on a given day and whether specific types of matches generated higher rates of success overall. We find strong evidence of the protocol effectively establishing close matches, and we also observe that matches based on education proved especially effective for rates of “success” in a panel survey that focused primarily on labor market topics. We conclude with thoughts on practical implementation of this approach in other settings and suggested directions for future work in this area.

KEYWORDS: Computer-assisted telephone interviewing; Interviewer effects; Interviewer-respondent interaction; Survey nonresponse.

1. INTRODUCTION

There is arguably no greater threat to the quality of modern survey research than unit nonresponse (Brick and Williams 2013; Williams and Brick 2017). This is especially true in telephone surveys, where response rates have declined substantially over the last decade (Curtin, Presser, and Singer 2005; Lavrakas, Benson, Blumberg, Buskirk, Flores, et al. 2017). Survey researchers need to develop innovative methods to mitigate the risk of nonresponse bias in survey estimates.

Interviewer administration is an important feature of the telephone survey that can help to maintain or improve response rates. Interviewers are responsible for establishing contact, introducing the survey, summarizing its objectives, addressing potential concerns, and securing cooperation in a limited window of interaction with a sampled person. Given this limited window, careful interviewer training and well-chosen design features are important for improving cooperation rates (Groves, Fowler, Couper, Lepkowski, Singer, et al. 2011). One common approach in the telephone survey context is to assign the highest-performing interviewers to the most difficult cases (which in many cases have expressed prior refusals), in hopes of eventually securing cooperation (Luiten and Schouten 2013; Schouten, Peytchev, and Wagner 2017, Chapter 5). This approach has been shown to increase response rates or produce more balanced response across characteristics available on the sampling frame relative to random assignment of interviewers to active sample cases.

An alternative approach that survey researchers may be able to employ, especially in the case of telephone panel surveys or surveys with rich sampling frame information, is matching the sociodemographic characteristics of sampled persons to interviewers prior to making contact attempts. A modest literature considers the effects of this type of respondent-interviewer matching, most of which focuses on face-to-face surveys. Brehm (1993, Chapter 3), Webster

(1996), and Moorman, Newman, Millikan, Tse, and Sandler (1999) each found that respondents contacted by interviewers of the same racial/ethnic background were more likely to cooperate with the survey request, and these findings were echoed in the telephone survey setting by Kim, Krosnick, and Lelkes (2018). Lord, Friday, and Brennan (2005) found evidence in support of matching interviewers with respondents on age and race, and both Durrant, Groves, Staetsky, and Steele (2010) and Vercruyssen, Wuyts, and Loosveldt (2017) reported higher cooperation rates when interviewers and respondents were aligned in terms of education and gender. Wang, Kott, and Moore (2013), however, reported that matching on race/ethnicity did not necessarily increase cooperation rates.

These conflicting findings could be due to poor understanding of the underlying causal mechanisms. Several influences on gaining respondent cooperation in interviewer-administered surveys have been discussed (Groves, Cialdini, and Couper 1992; West and Blom 2017). These include “liking” (if the interviewer is viewed favorably by the respondent) and “authority” (if the interviewer is perceived as either being or representing a credible and/or knowledgeable expert). These theories and principles may generate competing hypotheses in certain situations. For example, liking would suggest that an interviewer with the same sociodemographic characteristics as the respondent is more likely to gain cooperation. Participation decisions generally occur very quickly in telephone surveys—as quickly as the first few seconds of the interviewer-respondent interaction. This would suggest that many of the participation decisions are not based on thoughtful consideration of the purpose of the survey but instead on cognitive shortcuts (e.g., “I am like the person making the survey request, and this person finds the survey as important, so it must be important to me and I will cooperate.”). If so, this would further support the role of principles such as liking in survey participation decisions.

The authority principle suggests that interviewers perceived by respondents as being more credible and/or knowledgeable would be more likely to gain cooperation in general. For example, respondents may generally perceive older interviewers as having more knowledge on a variety of topics, and this perception of credibility may increase the likelihood of participation in surveys. Specific mismatches between sociodemographic characteristics may also engender perceptions of authority (e.g., older interviewers may be more effective at convincing younger persons to participate in a survey). Such causal links between interviewer and respondent characteristics and survey participation may be specific to a demographic characteristic (i.e., age, gender, education). Thus, matching might help for some characteristics but not for others.

Statistically, when modeling response propensity, one could detect the liking mechanism in the context of a specific sociodemographic characteristic (e.g., age) by testing an interaction between the interviewer characteristic and the respondent characteristic. This modeling approach has been used

Figure 1. Hypothetical Scenario Involving an Interaction Between Respondent Age and Interviewer (IWER) Age in the Prediction of Response Propensity, Where Matching on Age Tends to Be the Most Effective Strategy.

to test the plausibility of the liking mechanism in multiple prior studies (e.g., Webster 1996; Lord et al. 2005; Durrant et al. 2010). In the case where each possible characteristic has two possible values (e.g., older/younger), this approach would allow one to model the unique response propensity in all four cells of a two-by-two cross-classification of the two characteristics, and the liking mechanism would suggest that response propensity would be highest in the unique cells defined by matches (e.g., older respondent/older interviewer). The interaction would also enable detection of the scenario where the interviewer characteristic moderates the gap in response propensity between the groups defined by the respondent characteristic. For example, consider the scenario illustrated in figure 1: older respondents have a higher response propensity than younger respondents when the interviewer is older, and younger respondents have a higher response propensity than older respondents when the interviewer is younger.

Because including an interaction in the model enables one to model response propensity as a function of the cross-classification of the interviewer and respondent characteristics, this approach would also enable detection of a scenario where a *specific* type of match is more important for response

Figure 2. Hypothetical Scenario Involving an Interaction Between Respondent Age and Interviewer (IWER) Age in the Prediction of Response Propensity, Where Matching is Only Effective for Older Respondents.

propensity. For example, an older interviewer may optimize response propensity among older respondents, but the type of interviewer may not introduce a large difference in response propensity among younger interviewers (see [figure 2](#)). This is once again a scenario where the age of the interviewer is moderating the gap in response propensity between older and younger respondents and can be thought of as partial support for the liking mechanism.

Furthermore, if there is no evidence for the liking mechanism but the interaction is still significant, there may be support for the authority mechanism in the form of higher cooperation rates in mismatched cells (e.g., older interviewers are more likely to obtain cooperation from younger respondents). If the interaction is *not* significant, one could drop it from the model and evaluate main effects of either the interviewer or respondent characteristics on response propensity (see [figure 3](#), for example, where older respondents are more likely to respond in general). As noted earlier, the authority mechanism may also be supported if one were to find no evidence of an interaction but only an additive main effect of an interviewer characteristic that is likely to increase the respondent's perception of the interviewer as being credible and/or knowledgeable (e.g., age).

Figure 3. Hypothetical Scenario Where the Interaction Between Respondent Age and Interviewer (IWER) Age is Not Significant in Predicting Response Propensity, and There is No Support for the *Liking Mechanism* (Older Persons Are More Likely to Respond in General).

We note that one could also potentially analyze the liking mechanism by including a categorical predictor indicating the different combinations of socio-demographic characteristics for a given sampled person (e.g., older respondent/older interviewer, older respondent/younger interviewer, etc.), but this approach would not enable one to test the main effects associated with a particular characteristic (which could be used to support the authority mechanism).

There is a dearth of literature on the effectiveness of attempting to match the sociodemographic features of interviewers and respondents in the telephone survey context, presumably due to the absence of sufficient auxiliary information on sampled cases to perform the matching. Given the largely positive results in the small literature from face-to-face studies, a panel telephone survey would provide a convenient setting for studying the effectiveness of this kind of matching procedure experimentally. We therefore sought to implement a matching experiment in a panel survey employing computer-assisted telephone interviewing (CATI), focusing on those sociodemographic features identified as important in the matching literature. We sought answers to four research questions:

- (1) Does a protocol for attempting to match the sociodemographic features of interviewers and respondents based on the CATI interviewers available on a given day provide improvements in matching rates over standard CATI operations?
- (2) Does the matching protocol increase rates of success, defined as establishing a hard appointment or an interview at first contact?
- (3) Does the matching protocol reduce the time to a successful outcome?
- (4) What types of sociodemographic matches are most effective for increasing successful contacts in the panel telephone survey context?

2. METHODS

2.1 The PASS Study

We implemented this matching experiment in the eleventh wave of the PASS study, a German household panel survey for labor market, social policy, and poverty research with annual waves of data collection (Trappmann, Beste, Bethmann, and Müller 2013). The PASS questionnaire focuses on labor market participation, income sources, poverty, and social inclusion. The target population is the German residential population, with an oversampling of welfare benefit–recipient households. While respondents who once entered the panel remain in the sample as long as they are willing to participate, smaller refreshment samples from all new entries to welfare benefit receipt have been added every year. Additionally, there was one refreshment of the residential population sample before wave five.

The PASS study uses a sequential mixed-mode design combining CATI and computer-assisted personal interviewing (CAPI). The successful previous wave mode is the starting mode for the following wave. Through wave four, the default initial wave mode was CATI for all target households for whom a phone number was available, either from the sampling frame or from a directory search. Due to an increase in funding, the default initial wave mode for all refreshment samples from wave five on has been CAPI. The mode is switched if sample members cannot be contacted, refuse to participate, or express their wish to switch modes. In wave ten, about one-third of the interviews were conducted using CATI, leading to about the same distribution in starting modes for wave eleven.

Due to the specific panel design with yearly refreshments, the panel sample for wave eleven consists of households with a varying degree of panel experience. While more than one-third of the sample is from the original wave one sample and has participated for a full ten years, the remaining part of the sample has between one and nine years of experience with the panel. Some of these respondents may have missed one or more waves during the ten years, and temporary dropouts remain part of the panel sample unless they have missed two consecutive waves.

The experimental design was implemented for all existing panel households that started in CATI during wave eleven of PASS data collection. It was limited to the first phase of fieldwork (nine weeks) before any mode switches were implemented and to the household interview for which the target person is fixed and known in advance from the previous wave. Only households interviewed in German in the previous wave were included in the experiment described here. This leaves 2,965 households, out of which 2,520 (85.0 percent) were contacted, and 1,823 (61.5 percent, per AAPOR RR1) were interviewed during the first phase of fieldwork. We remind readers that this response rate only comes from phase one of the field work, prior to mode switching, and is therefore a conservative estimate/lower bound for what would have been the AAPOR RR1 calculation after all of the field work was completed.

2.2 Experimental Design

In an ideal setting with no operational or resource constraints, an experiment to test the effectiveness of the type of matching protocol under investigation here would randomize respondents to either always be attempted by an interviewer with matching sociodemographic characteristics or to always be attempted by the first available interviewer working a specified calling shift within a telephone center. This ideal design cannot be implemented in practice due to the substantial inefficiencies it would create: experimental cases would be delayed due to a lack of matching interviewers available at a given shift, and the management of two separate projects would become increasingly complex.

We therefore implemented an approximation of the ideal design, using an alternating “on/off” matching protocol. On the first day of data collection for wave eleven, the matching protocol was “off,” and active panel household cases using CATI were assigned to the first interviewer that was available to make a call. On the second day of data collection, the matching protocol was “on,” and the CATI system attempted to assign active panel households to available interviewers using a priority-based matching protocol (described in more detail later on). The alternating pattern of “off/on” matching protocols continued for the duration of the first phase of wave eleven data collection. This design ensured that the matching protocol was not restricted to specific calling shifts or to respondents who would only be available on particular days.

We now describe the matching protocol that was employed for “on” days. We defined eight sociodemographic subgroups of both interviewers and respondents, based on known values of gender, age, and education (see [table 1](#)). When a given panel household was available for calling, the sociodemographic features of the household respondent from the prior wave were extracted, and the system attempted to assign the case to an available interviewer according to the priorities indicated in [table 2](#). For example, an interviewer belonging to

Table 1. Subgroups of Interviewers and Respondents Based on Gender, Age, and Education

Subgroup	Gender	Age	Education
1	Male	< 40 years	No Advanced Education
2	Male	< 40 years	University / Vocational Education
3	Male	≥ 40 years	No Advanced Education
4	Male	≥ 40 years	University / Vocational Education
5	Female	< 40 years	No Advanced Education
6	Female	< 40 years	University / Vocational Education
7	Female	≥ 40 years	No Advanced Education
8	Female	≥ 40 years	University / Vocational Education

Table 2. Matching Protocol Indicating Matching Priorities Employed by the CATI System During “on” Days

Interviewer subgroup	Respondent match priority
1	1, 2, 3, 5, 4, 6, 7, 8
2	2, 1, 4, 6, 3, 5, 8, 7
3	3, 4, 1, 7, 2, 8, 5, 6
4	4, 3, 2, 8, 1, 7, 6, 5
5	5, 6, 7, 1, 8, 2, 3, 4
6	6, 5, 8, 2, 7, 1, 4, 3
7	7, 8, 5, 3, 6, 4, 1, 2
8	8, 7, 6, 4, 5, 3, 2, 1

subgroup one would be assigned a panel household with the most recent respondent belonging to subgroup one as well (the highest priority case), if possible. If such a respondent was not available, the system attempted to find a panel household with a primary respondent belonging to subgroup two, and so forth.

This matching scheme maximized the number of characteristics matched on; when only one or two characteristics could be matched, the matching priorities were gender first, then age, and finally education. The supervisors of the infas telephone center (where infas is the German data collection organization currently operating the PASS study) agreed with the use of these matching priorities, based on three factors: 1) the authors' intuition about the relative importance of each characteristic in influencing responding behavior (e.g., gender and age would be more immediately recognizable over the telephone than education); 2) the experience of infas with the PASS study from previous waves; and 3) discussion of the relevant literature (although as noted previously,

conflicting findings in this literature made it difficult to defend *a priori* one priority over another). This design is operationally feasible since it does not require that cases be matched to interviewers in order for attempts to be made, as discussed earlier.

2.3 Characteristics of Interviewers and Respondents Attempted Using CATI in Wave Eleven

There were 131 available telephone lab interviewers that attempted panel households using CATI in the first phase of wave eleven. Of these, sixty (45.8 percent) were male; sixty-two (47.3 percent) were younger than forty years old; and eighty-six (65.6 percent) had a vocational or university education. Among the 2,965 respondents in the panel households that were attempted using CATI in the first phase of wave eleven, 1,315 (44.4 percent) were male; 703 (23.7 percent) were younger than forty years old; and 2,458 (82.9 percent) had a vocational or university education. We note that in Germany, having a university degree or a vocational education (in most cases in a dual system, combining apprenticeships in a company and vocational education at a vocational school in one course) is a prerequisite for many occupations. The available panel respondents were, therefore, older and more educated than the interviewers in general.

2.4 Analytic Approach

In all analyses, standard errors and significance tests were adjusted to account for clustering of observations within interviewers (Donner and Klar 2000). A “successful outcome” was defined as establishing a hard appointment or completing an interview.

For our first research question, we examined the proportion of matches by priority (first priority, second priority, etc.) for all attempted calls on “on” days and “off” days and compared these distributions statistically using a χ^2 test. Specifically, we used the user-written “clchi2” command in Stata to adjust these χ^2 tests for clustering by interviewer (Herrin 2002). We also compared matching rates for each sociodemographic characteristic separately on “on” and “off” days, again using χ^2 tests. We then repeated these analyses for *first contacts only*. For our second research question, we used a χ^2 test to compare the probability of success between “on” and “off” days, restricted to first contacts.

For our third research question, we used a discrete-time hazard modeling approach to examine the number of contacts until successful contact, with cases that did not have a successful outcome considered as right-censored. This model included an interaction between contact number and daily protocol status to assess whether the protocol was more effective for earlier contacts.

For our fourth research question, we first examined proportions of successful outcomes at first contact depending on the type of match (i.e., gender only, age only, education only, gender/age, age/education, gender/education, or all three characteristics) and compared the proportions statistically using the aforementioned cluster-adjusted χ^2 test. We then fitted logistic regression models predicting a successful outcome at first contact with two-way interactions between selected respondent and interviewer characteristics (e.g., interviewer gender $\times$ respondent gender), controlling for the number of days since the beginning of the data collection, daily protocol status, an indicator of whether the contact occurred on a weekend, an interviewer-specific measure of the number of years since their first PASS interview was completed (as a proxy of interviewer experience), and the number of years since a respondent completed their first PASS interview (as a proxy of experience with the panel). Specifically, we used the `logit` command in Stata to fit the logistic regression models, and we used the “`vce`” (`cluster INTID`) option to adjust the standard errors of the estimated regression parameters for the clustering of the data by interviewer ID.

Three initial models considered one characteristic only (e.g., gender), and a subsequent model considered all three characteristics simultaneously. The two-way interactions between the characteristics were removed from the models if they were not found to be significant (suggesting no support for the liking mechanism). After fitting each final model, we used the `margins` post-estimation command to compute marginal model-based predicted values for the probability of success at first contact for each cell defined by the 2×2 cross-classification of a given respondent characteristic and an interviewer characteristic. These were the average predicted probabilities for all cases included in each final model, assuming that all cases belonged to a given cell and incorporating the values of all other covariates. Standard errors for these marginal predicted values were computed based on the delta method. All analyses were performed using Stata (Version 15.1), and examples of the code used and output generated can be found in the [Supplementary Materials](#).

3. RESULTS

3.1 Research Question 1: Was the Matching Protocol Effective?

There was a substantial shift in the probability of a highest priority match between interviewer and respondent on the “on” days when the matching protocol was implemented: 56.3 percent of calls on “on” days resulted in a highest priority match (per [table 2](#)) versus 17.7 percent of calls on “off” days. Nearly 94 percent of calls on “on” days resulted in a match of the top four priorities, compared with less than 58 percent of calls on “off” days. Considering the specific sociodemographic characteristics, 91.1 percent of calls on “on” days were

gender matched, versus 48.8 percent of calls on “off” days; 76.4 percent of calls on “on” days were age matched, versus 50.6 percent of calls on “off” days; and 79.5 percent of calls on “on” days were education matched versus 63.0 percent of calls on “off” days. We note that the overall “highest priority” match rate on the “off” days of 17.7 percent approximates the rate that would be expected under independent random matching of the three factors ($0.488 \times 0.506 \times 0.630 = 0.156$).

When considering first contacts only, 75.2 percent of first contacts on “on” days resulted in the highest-priority match versus only 20.2 percent of first contacts on “off” days. Very similar differences in matching rates were observed when considering the specific sociodemographic matches and first contacts only. All of the comparisons reported for research question one had corresponding χ^2 statistics that were significant at the 0.001 level. Collectively, these results suggest that the matching protocol as described previously was highly effective at engendering the matches between respondents and interviewers based on all three sociodemographic characteristics.

3.2 Research Question 2: Were There Higher Rates of Success at First Contact on Days When the Matching Protocol Was Implemented?

Overall, when considering the 2,520 first contacts with panel households during the course of the experiment, 57.1 percent of those first contacts occurring on “on” days ($n = 1,382$) were successful (i.e., ended with a hard appointment or a complete interview) versus 56.4 percent of first contacts occurring on “off” days ($n = 1,138$). This difference in success rates was slightly in favor of the matching protocol but not statistically significant.

3.3 Research Question 3: Did the Matching Protocol Affect the Time to a Successful Outcome?

We did not find the interaction between contact number and “matching” day to be significant in the discrete time hazard model. When removing the nonsignificant interaction, the main effect of “matching” day (“on” versus “off”) was also not significant. In short, there was no effect of an “on” day contact on the “hazard” of a successful outcome, and this did not change depending on the contact number. We note that there were only 1.7 contacts per household on average.

3.4 Research Question 4: What Types of Matches Were Most Effective for Success at First Contact?

While research question two considers the overall effect of the matching protocol, it does not consider the specific nature of the match. We first consider

Table 3. Percentages of First Contacts with Successful Outcomes by Respondent-Interviewer Dyads Based on Gender

Respondent-interviewer dyad	Percentage of first contacts resulting in completed interview or hard appointment (Success)
Male-Male	54.62% ($n = 910$)
Male-Female	56.82% ($n = 220$)
Female-Male	51.69% ($n = 414$)
Female-Female	60.96% ($n = 976$)
χ^2 test	$\chi^2 (3) = 13.07, p = 0.004$

descriptive results, examining the ability of each type of match to increase rates of success at first contact. Table 3 shows percentages of first contacts with successful outcomes by respondent-interviewer combinations based on gender and suggests that female interviewers tended to garner more success regardless of whether the respondent was male or female. There is little evidence here to support the liking mechanism, but we revisit this result after fitting our regression models.

Table 4 provides slightly more evidence in favor of the liking mechanism, where younger interviewers improve the success rate by about four percentage points for younger respondents, and older interviewers have a similar effect for older respondents.

Table 5 provides the strongest support for the liking mechanism, especially for respondents with lower education. Among this subgroup, using an interviewer with lower education increased the success rate by more than fifteen percentage points; among the subgroup of respondents with higher education, using an interviewer with higher education increased the success rate by about 2.5 percentage points.

Finally, table 6 presents descriptive success rates based on the overall quality of the match, considering all three characteristics simultaneously. While these results do not adjust for any potential confounding factors and actual matches did not occur at random but rather depended on both the availability of appropriate interviewers and on our prioritization of matches, there is a general suggestion that matches involving education and matches based on all three characteristics tended to engender the highest success rates.

We now consider our multivariable results and determine whether the descriptive results presented previously remain robust when adjusting for several other potential confounding factors. Table 7 presents results from the four separate logistic regression models fitted to the indicator of success at first contact.

In the first model, focusing on gender, we did not find significant evidence of an interaction between respondent gender and interviewer gender, and this interaction was thus dropped from the model. No variables outside of respondent experience on the panel were found to be significant predictors of success

Table 4. Percentages of First Contacts with Successful Outcomes by Respondent-Interviewer Dyads Based on Age

Respondent-interviewer dyad	Percentage of first contacts resulting in completed interview or hard appointment (Success)
Young-Young	48.19% ($n = 415$)
Young-Old	43.82% ($n = 89$)
Old-Young	55.98% ($n = 661$)
Old-Old	60.66% ($n = 1,355$)
Chi-square test	$\chi^2 (3) = 27.07, p < 0.001$

Table 5. Percentages of First Contacts with Successful Outcomes by Respondent-Interviewer Dyads Based on Education

Respondent-interviewer dyad	Percentage of first contacts resulting in completed interview or hard appointment (Success)
Low-Low	53.71% ($n = 229$)
Low-High	38.26% ($n = 149$)
High-Low	56.11% ($n = 360$)
High-High	58.87% ($n = 1,782$)
Chi-square test	$\chi^2 (3) = 24.94, p < 0.001$

Table 6. Percentages of First Contacts with Successful Outcomes by Quality of Match, Considering all Three Characteristics Simultaneously

Type of match	Percentage of first contacts resulting in completed interview or hard appointment (success)
None	55.93% ($n = 118$)
Sex Only	53.66% ($n = 164$)
Age Only	45.71% ($n = 70$)
Education Only	50.29% ($n = 171$)
Sex and Age	47.13% ($n = 157$)
Sex and Education	57.24% ($n = 297$)
Age and Education	57.45% ($n = 158$)
All Three Characteristics	60.49% ($n = 1,268$)
Chi-square test	$\chi^2 (7) = 20.13, p < 0.01$

in this initial model. In the second model, focusing on age, we again found no evidence of an interaction between respondent age and interviewer age. These results do not provide support for the liking mechanism based on gender or

Table 7. Estimated Coefficients (with Robust Standard Errors in Parentheses, Accounting for Within-Interviewer Correlation in Success) in Logistic Regression Models Predicting Success at First Contact as a Function of Sociodemographic Features of Respondents and Interviewers

Predictor	Model 1: Sex	Model 2: Age	Model 3: Education	Model 4: Overall
Intercept	−0.351 (0.249)	−0.400 (0.238)	−0.166 (0.235)	−0.289 (0.254)
R Female	0.052 (0.103)			0.066 (0.098)
I Female	0.146 (0.181)			0.133 (0.169)
R ≥ 40 years old		0.268 (0.162)		0.214 (0.155)
I ≥ 40 years old		0.053 (0.265)		0.095 (0.268)
R University or Vocational Educ.			0.001 (0.191)	−0.008 (0.179)
I University or Vocational Educ.			−0.771 (0.255)***	−0.788 (0.237)***
R University / Vocational Educ. × I University / Vocational Educ.			0.792 (0.242)***	0.773 (0.237)***
Matching “On”				
Day	0.012 (0.121)	0.043 (0.122)	−0.076 (0.124)	−0.035 (0.119)
Weekend	−0.014 (0.008)	−0.015 (0.008)	−0.014 (0.008)	−0.015 (0.008)
I Years Since First PASS Interview	−0.163 (0.225)	−0.104 (0.233)	−0.159 (0.220)	−0.140 (0.226)
R Years Since First PASS Interview	0.026 (0.034)	0.023 (0.048)	0.037 (0.035)	0.017 (0.044)
R Years Since First PASS Interview	0.088 (0.018)***	0.077 (0.018)***	0.084 (0.019)***	0.070 (0.017)***
n	2,520	2,520	2,520	2,520
Log-likelihood	−1685.789	−1684.400	−1678.241	−1673.034
Pseudo R2	0.020	0.021	0.025	0.028

***p ≤ 0.01, **p ≤ 0.05; I = Interviewer, R = Respondent, Day = Days since start of data collection; Pseudo R² is using the calculation proposed by [McFadden \(1973\)](#).

age. When considering the authority mechanism, one could speculate that respondents may have perceived older interviewers as having more knowledge about the labor market in general, meaning that older interviewers would be perceived as representing authority on the topic and thus more likely to obtain cooperation. However, given that there were also no main effects of interviewer age or gender after dropping the nonsignificant interactions, we did not find any support for the authority mechanism either.

In the third model, we found strong evidence of an interaction between respondent education and interviewer education, and these results remained robust in the fourth model controlling for the age and gender of both the respondent and the interviewer. [Table 8](#) presents marginal predicted probabilities of success at first contact based on the fourth model in [table 7](#), in addition to standard errors for the marginal predictions. Largely consistent with the descriptive results in [table 5](#), we see that matching lower education respondents with lower education interviewers is much more productive than matching lower education respondents with higher education interviewers. However, we do not see the same benefit of matching based on education among respondents with higher education, where the education of the interviewer essentially does not matter. We therefore find partial support for the liking mechanism and no support for the authority mechanism, and the scenario realized here is similar to that which was illustrated in [figure 2](#).

4. DISCUSSION

4.1 Summary of Results

This study successfully implemented a matching protocol in a CATI environment that aligned sociodemographic features of interviewers with those of respondents, using a long-running panel study in Germany. While we did not conduct a true randomized experiment, we demonstrated that alternating days of having the matching protocol “on” and then having it “off” was able to substantially increase matching rates on the “on” days. When considering the establishment of a hard appointment for a future interview or a completed interview as a success, the rate of success was only slightly higher on the “on” days when the matching protocol was successfully implemented (considering first contacts). The protocol also did not decrease the amount of time until a successful outcome.

However, we did find that specific types of matches, irrespective of their experimental manipulation, were effective for increasing rates of success at first contact. Specifically, attempting to gain cooperation using an interviewer with lower education for a respondent with lower education was successful. These results are consistent with those of [Durrant et al. \(2010\)](#) and [Vercruyssen et al. \(2017\)](#) in face-to-face surveys and provide important new evidence about the

Table 8. Marginal Predicted Probabilities of Success at First Contact Based on Model 4 in Table 7, by Respondent-Interviewer Dyads Based on Education

Respondent-interviewer dyad	Marginal predicted probability of first contact resulting in completed interview or hard appointment (Delta method SE)
Low-Low	0.587 (0.036)
Low-High	0.398 (0.045)
High-Low	0.585 (0.032)
High-High	0.582 (0.028)

effectiveness of this type of matching in the CATI setting. They also provide further evidence of people resorting to cognitive shortcuts and heuristics when making participation decisions. Panel members with lower education who have been interviewed multiple times about their job status (or receipt of unemployment benefits) know that they will be asked about these topics, and this has the potential to exacerbate negative feelings about their economic position. Speaking with an interviewer perceived to have higher social status, whatever the heuristic is that leads to that perception, may increase the strength of these feelings, unlike panel members with higher education (and potentially higher socio-economic status). This could ultimately lead to nonparticipation, despite being a panel member.

So what heuristics might respondents have been using to predict the education status of the interviewer in the CATI setting? While all PASS interviewers are trained to follow the same interviewing protocol, it may be the case that interviewers with higher and lower education levels differ from each other in their approach to gaining participation or they might adjust their approach based on what they can perceive about the respondent. For example, interviewers with higher education might pick up on respondent concerns or difficulties with the survey request more quickly, follow the script more strictly (appearing to be more “robotic”), speak more quickly (or with greater diction), or attempt to empathize more with the respondent’s stated concerns. All of these actions may be interpreted differently by respondents with varying levels of education, ultimately leading to changing likelihoods of participation.

The respondents may have also been judging the education of the interviewers based on *how* they were speaking. In everyday conversations, the language used is characterized by different cultural markers, including accents, regional dialects, formality of speech, and pronunciation. Respondents may have been judging the interviewer’s socioeconomic status based on subtle cues present in the language that they were using. In this case, the causal mechanism underlying the effectiveness of interviewer-respondent matching based on education would be independent of any conscious action or particular utterance of the interviewer. Even in the brief window of time that one is invited to

participate in a survey over the telephone, the low-education respondent may have simply detected linguistic cues indicating higher or lower socioeconomic status of the interviewer and reacted accordingly. While we did not have sufficient information from this study to test this potential mechanism more formally, there is literature suggesting that US respondents can identify the racial background of an interviewer over the telephone (Davis 1997; Davis and Silver 2003), and future research could explore this potential mechanism further.

Whether or not these mechanisms are reasons for the effectiveness of the education matching in this study is an open question that warrants future research. In the absence of empirical information about these behaviors, possibly based on a large sample of audio recordings of the interactions, it is hard to make formal training suggestions based on our speculation (outside of a general suggestion to try to match interviewers and respondents based on education). Given that similar findings related to education matching were also reported by Durrant et al. (2010) and Vercruyssen et al. (2017), more research into the underlying mechanisms is warranted.

The audio recording of survey introductions is not permitted in Germany. Recordings only start after the respondent has given permission to record the interview (and this question is only asked at the start of the interview, *after* cooperation has been achieved). Interviewer surveys, interviewer debriefings, or live monitoring of telephone introductions would be the only sources of more information about these potential mechanisms in future CATI studies. Along these lines, the PASS study has conducted interviewer surveys in prior waves. We analyzed data from an interviewer survey that was conducted before wave five, and these analyses suggested that it is unlikely that interviewers with different educational degrees differ in their recruitment strategies. For seven of the eight survey items related to the attitudes of the interviewers toward gaining cooperation, there were no significant differences in response distributions between interviewers with university degrees or vocational education and interviewers without. The only difference that we found was for the item, “One should always emphasize the voluntary nature of participation,” where interviewers with advanced education were more likely to agree.

It may also be the case that a stronger impact of matching would be obtained in a different type of survey. The experiment reported here was conducted on the eleventh wave of a panel survey. Panel members had much more information about the survey than just the interviewer’s voice and the first few words that they used. Further, due to panel attrition over the course of ten previous waves, the group eligible for the eleventh wave was also likely to be more cooperative than a random sample of the population. This factor—a highly cooperative group that is also very familiar with the survey—may have limited the effect of the experimental matching protocol. Replications with fresh cross-

sectional samples of the population drawn from frames including rich auxiliary information may produce stronger results.

4.2 Concrete Suggestions for Practice

In a one-time CATI survey, initial contact attempts could yield information about interviewer perceptions of respondent features, such as gender (McCulloch, Kreuter, and Calvano 2010), and this information could be used to establish matches at future contact attempts. Alternatively, in the absence of any prior auxiliary information on sampled telephone numbers or households, linking commercial data to the sampled telephone numbers might prove to be effective (West, Wagner, Hubbard, and Gu 2015). For panel surveys employing CATI, we would suggest replicating the study considered here to first identify the most effective types of matches in a particular survey context.

4.3 Additional Directions for Future Research

We were unable to consider other sociodemographic characteristics that may be useful for matching, such as race, ethnicity, and dialect; such studies would be helpful next steps. We also note that the success by type of match observed in this study was dependent on the protocol that we used, which matched sociodemographic characteristics according to a predefined priority (sex, then age, then education). Hence, the ability to detect the effectiveness of matching by education was diminished, given that matches based on education tended to be conditional on first matching by sex and then by age. This could be a reason for the slight positive effect of being contacted on an “on” day on overall rates of success; if education had been the *first* priority, the success rate on the “on” days would likely have been much higher. This underscores the need for future studies to first understand what the most effective types of matches will be before attempting to implement the matching protocol studied here. Additional evaluations of the effectiveness of matching in the cross-sectional CATI context and surveys with lower response rates would also be valuable, as effects may be much larger and possibly different depending on context.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

References

- Brehm, J. O. (1993), *The Phantom Respondents: Opinion Surveys and Political Representation*, Ann Arbor, MI: University of Michigan Press.

- Brick, J. M., and D. Williams (2013), "Explaining Rising Nonresponse Rates in Cross-sectional Surveys," *The Annals of the American Academy of Political and Social Science*, 645, 36–59.
- Curtin, R., S. Presser, and E. Singer (2005), "Changes in Telephone Survey Nonresponse Over the Past Quarter Century," *Public Opinion Quarterly*, 69, 87–98.
- Davis, D. W. (1997), "Nonrandom Measurement Error and Race of Interviewer Effects Among African Americans," *Public Opinion Quarterly*, 61, 183–207.
- Davis, D. W., and B. D. Silver (2003), "Stereotype Threat and Race of Interviewer Effects in a Survey on Political Knowledge," *American Journal of Political Science*, 47, 33–45.
- Donner, A., and N. Klar (2000), *Design and Analysis of Cluster Randomization Trials in Health Research*, London: Arnold.
- Durrant, G. B., R. M. Groves, L. Staetsky, and F. Steele (2010), "Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys," *Public Opinion Quarterly*, 74, 1–36.
- Groves, R. M., R. B. Cialdini, and M. P. Couper (1992), "Understanding the Decision to Participate in a Survey," *Public Opinion Quarterly*, 56, 475–495.
- Groves, R. M., F. J. Fowler, Jr, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2011), *Survey Methodology* (2nd ed.), New York: Wiley.
- Herrin, J. (2002), "CLTEST: Stata modules for performing cluster-adjusted chi-square and t-tests," Statistical Software Components S424901, Boston College Department of Economics, last revised 03 Feb 2012.
- Kim, N., J. A. Krosnick, and Y. Lelkes (2018), "Race of Interviewer Effects in Telephone Surveys Preceding the 2008 U.S. Presidential Election," *International Journal of Public Opinion Research*, DOI: 10.1093/ijpor/edy005.
- Lavrakas, P. J., G. Benson, S. Blumberg, T. Buskirk, I. Flores, L. Christian, D. Dutwin, M. Fahimi, H. Fienberg, T. Guterbock, S. Keeter, J. Kelly, C. Kennedy, A. Peytchev, L. Piekarski, and C. Shuttles (2017), "AAPOR Task Force Report on The Future of U.S. General Population Telephone Survey Research," Available at <http://www.aapor.org/Education-Resources/Reports/The-Future-Of-U-S-General-Population-Telephone-Sur.aspx>, Accessed August 29, 2017.
- Lord, V. B., P. C. Friday, and P. K. Brennan (2005), "The Effects of Interviewer Characteristics on Arrestee's Responses to Drug-related Questions," *Applied Psychology in Criminal Justice*, 1, 36–55.
- Luiten, A., and B. Schouten (2013), "Tailored Fieldwork Design to Increase Representative Household Survey Response: An Experiment in the Survey of Consumer Satisfaction," *Journal of the Royal Statistical Society*, 176, 169–189.
- McCulloch, S. K., F. Kreuter, and S. Calvano (2010), "Interviewer Observed vs. Reported Respondent Gender: Implications on Measurement Error," Paper presented at the 2010 Annual Meeting of the American Association for Public Opinion Research, Chicago, IL.
- McFadden, D. (1973), "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Economics*, ed. P. Zarembka, New York, NY, Academic Press, p. 105–142.
- Moorman, P. G., B. Newman, R. C. Millikan, C. J. Tse, and D. P. Sandler (1999), "Participation Rates in a Case-Control Study: The Impact of Age, Race, and Race of Interviewer," *Annals of Epidemiology*, 9, 188–195.
- Schouten, B., A. Peytchev, and J. Wagner (2017), *Adaptive Survey Design*, Boca Raton, FL: CRC Press.
- Trappmann, M., J. Beste, A. Bethmann, and G. Müller (2013), "The PASS Panel Survey after Six waves," *Journal for Labour Market Research*, 46, 275–281.
- Vercruyssen, A., C. Wuyts, and G. Loosveldt (2017), "The Effect of Sociodemographic (Mis) Match Between Interviewers and Respondents on Unit and Item Nonresponse In Belgium," *Social Science Research*, 67, 229–238.
- Wang, K., P. Kott, and A. Moore (2013), "Assessing the Relationship Between Interviewer Effects and NSDUH Data Quality," Report prepared by Research Triangle Institute for the Substance Abuse and Mental Health Services Administration, Research Triangle Park, NC.
- Webster, C. (1996), "Hispanic and Anglo Interviewer and Respondent Ethnicity and Gender: The Impact on Survey Response Quality," *Journal of Marketing Research*, 33, 62–72.
- West, B. T., and A. G. Blom (2017), "Explaining Interviewer Effects: A Research Synthesis," *Journal of Survey Statistics and Methodology*, 5, 175–211.

- West, B. T., J. Wagner, F. Hubbard, and H. Gu (2015), "The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth," *Journal of Survey Statistics and Methodology*, 3, 240–264.
- Williams, D., and J. M. Brick (2017), "Trends in US Face-to-Face Household Survey Nonresponse and Level of Effort," *Journal of Survey Statistics and Methodology*, 6, 186–211.