The PAIP Score: A Propensity-Adjusted Interviewer Performance Indicator

Brady T. West[1], Robert M. Groves[2]

[1] Michigan Program in Survey Methodology, Institute for Social Research, 426 Thompson Street, Ann Arbor, MI, 48104

[2] U.S. Census Bureau, 4600 Silver Hill Rd., Washington, D.C., 20233

Abstract

Evaluating interviewers based on their response rates is complicated in most surveys. By random chance, interviewers may call cases that are more or less difficult to interview. In addition, interviewer response rates can only imperfectly be computed because of the contributions of other interviewers' prior contacts with those cases (calling a case after a contact from an expert interviewer may pose different difficulties than calling a case after a contact from an inexperienced interviewer). This paper proposes and evaluates an interviewer performance indicator that attempts to repair these weaknesses. The proposed indicator is computed using a three-step algorithm. First, for each active case, available paradata are used to estimate the propensity that the next contact with the case will generate an interview. Second, if the interviewer assigned the case obtains a successful interview on the next contact, the interviewer receives a score of 1 minus the estimated response propensity for the contact; a non-successful contact by the interviewer results in a score of 0 minus the estimated response propensity for the contact. Finally, for each interviewer, the contact-level scores are averaged over all contacts, resulting in a propensity-adjusted interviewer performance (PAIP) score. Addressing an important drawback of previous interviewer performance measures discussed in the literature, this performance indicator gives large credit to the interviewer who obtains success on very difficult cases, and only a small penalty given failure with such cases. The indicator gives only small credit to success on very easy cases and larger penalties given failure with easy cases. This paper illustrates computation of the PAIP score using two different surveys (one face-to-face, one telephone), and assesses the validity of the indicator as a new metric for evaluating the performance of interviewers.

Key Words: Interviewer Evaluation, Interviewer Performance, Response Propensity Modeling

1. Introduction

Interviewer-level cooperation rates are often used as evaluative tools of interviewer performance (e.g., Fowler, 2008; Groves and Couper, 1998; Kennickell, 2006; Tarnai and Moore, 2008). In area probability sample surveys, interpreting such rates is complicated. Interviewers whose sample assignments lie in large urban areas tend to obtain lower cooperation rates than those working in small rural communities (Groves and Couper, 1998). These differences are believed to arise because of either target population differences that are out of the control of the survey organization (such as urbanicity) or interviewer differences (e.g., Singer et al., 1983), and a large body of literature provides empirical support for the existence of between-interviewer variance in cooperation rates in face-to-face surveys (Hox and de Leeuw, 2002; Morton-Williams, 1993; O'Muircheartaigh and Campanelli, 1999; Snijkers, Hox, and de Leeuw, 1999; Wiggins, Longford, and O'Muircheartaigh, 1992; but see Hox, de Leeuw and Kreft, 1991). In contrast, evaluation of interviewers in centralized telephone interviewing facilities is facilitated by the fact that interviewers share sample cases over time, thus yielding workloads of more homogeneous difficulty, especially within shifts. However, even in centralized telephone interviewing, there remain three complications of interpreting interviewer-level cooperation rates:

a) Interviewers by chance call cases that are more or less difficult to interview, and interviewers assigned refusal conversion cases (primarily in telephone surveys) have greater (but usually unmeasured) burden to obtain a given response rate;
b) Interviewer response rates can only imperfectly be computed because of the contributions of other interviewers' prior contacts with those cases (calling a case after a contact from an expert interviewer in a telephone survey may pose different difficulties than calling a case after a contact from a less-skilled interviewer); and
c) Evaluations of interviewers typically are based on case outcomes; it could be argued that the real performance unit of an interviewer concerns the outcome of a contact, not a case.

For these reasons, it would be useful to the survey practitioner to have a metric for comparing the effectiveness of interviewers that incorporates *difficulty* of the interviewing task at the contact level. Much of the survey literature focuses on measures for managing interviewer productivity, including number of completed interviews, calls attempted per hour, hours per completed interview, or refusals per hour (Tarnai and Moore, 2008). Productivity metrics are certainly important for managing the timeliness of existing survey operations and projecting future budgets. However, metrics for evaluating interviewer performance (e.g., calls and completes per hour, measures of data quality, refusal rates, etc.) are often comparative in nature, either based on past performance or the current performance of other interviewers (Barioux, 1952; Tarnai and Moore, 2008), and do not recognize the difficulty of the assigned cases in the current study. For example, an interviewer may have performed very well in previous surveys according to a variety of metrics, but if she is assigned a very difficult set of cases in the current survey, her evaluation measures may suffer in comparison.

The various measures of interviewer performance that have been proposed in the literature (e.g., Abbott et al., 2004; Anton, 1997; Peng and Feld, 2011; Summers and Beck, 1973; Waite, 2002) generally fail to incorporate the difficulty of assigned cases. For example, Berk and Bernstein (1988) used item nonresponse rates and response validity based on a follow-up verification survey to study the associations between performance and interviewer education and experience. However, recent work has shown that more difficult cases may also produce more measurement error, depending on the statistic of interest (Olson, 2006). Sheatsley (1949) examined changes in the proportion of interviewers accepting "Yes" responses that would lead to a longer subset of questions across surveys, but this evaluation measure was an aggregate measure rather than an interviewer-specific measure. Manheimer and Hyman (1949) used accuracy in the listing of dwelling units, selection of samples of dwelling units, and selection of individuals within dwelling units to evaluate interviewer performance, but these measures all require validation data and may vary depending on the difficulty (e.g., geographic complexity) of the area being worked.

Sudman (1966) proposed evaluating interviewer performance across studies by adjusting for the difficulty of a *study*, but did not discuss variance in the difficulty of cases *within* a study. Durand (2005) developed a scoring system for interviewer performance in a telephone survey (the Net Contribution to Performance Index, or NCPI) designed to award more points for converted refusals, thus recognizing the type of case that an interviewer was working (never-reached, previous refusal, or appointment). Durand (2008) further illustrated several advantages of the NCPI over the more commonly used cooperation rate at first contact (COOPRT1) performance metric. Laflamme and St. Jean (2011) proposed a weighted interviewer performance (WIP) measure for CATI surveys based on current survey productivity in cells defined by three factors: the type of case being worked (e.g., a contacted case without any prior refusals), the amount of work already performed on a case, and the time of day at which the call was made. This methodology, used for CATI surveys by Statistics Canada, partially incorporates the difficulty of cases currently being worked, but is restrictive in that only three factors are used to define the difficulty of a case. Indeed, the authors state that "…any WIP measure needs to be analyzed in its environmental context. A lower WIP does not necessarily mean that the interviewer is not as good as others. It can simply mean that the

interviewer has worked on more difficult cases and this information needs to be provided and considered by the RO survey manager in the evaluation" (Laflamme and St. Jean, 2011, 7). In this paper, we build on these important prior efforts to develop performance metrics that incorporate case difficulty, and propose a more general method for evaluating interviewer performance that incorporates a continuous measure of call difficulty and can be applied in CATI or CAPI surveys.

Given the general rise in refusals to participate in surveys (e.g., Curtin et al., 2005; de Leeuw and de Heer, 2002), it might be useful to focus an interviewer performance indicator on the ability of an interviewer to obtain an interview, given successful contact with a sample case. In this paper, we propose a metric that can be used to evaluate interviewers which focuses on successful contacts with sample cases, and incorporates the projected difficulty of those cases based on response propensity modeling. This extends the ideas proposed by Durand (2005) and Laflamme and St. Jean (2011) to allow for differences in expected response propensity between cases with similar classifications (e.g., two previous refusals); interviewers successfully obtaining cooperation from a more difficult previous refusal (based on expected response propensity) would be given a larger score for the contact. We focus on ability to obtain an interview upon contact, rather than performance during the interview; see Biemer et al. (2000), Cannell et al. (1975), Edwards et al. (1994) or Herget et al. (2001) for recording strategies that have been used to evaluate performance during the interview. Although overall evaluations of interviewers should certainly be based on multiple metrics (Steinkamp, 1964), the metric proposed in this paper provides survey managers with a contact-based measure of ability to obtain interviews that adjusts for a continuous measure of the difficulty of assigned cases.

2. A Propensity-Adjusted Interviewer Performance Metric: The PAIP Score

With appropriate survey paradata (Couper, 1998; Couper and Lyberg, 2005), an interviewer performance indicator might be considered that measures the interviewer's average performance in obtaining interviews over all the contacts made with sample households. The use of paradata, or "data collection process" data, to estimate response propensities and possibly repair nonresponse bias has received a considerable amount of recent research attention (e.g., Beaumont, 2005; Chun et al., 2010; Jocelyn and Baribeau, 2010; Kreuter et al., 2010; Stoop et al., 2010). In this paper, we build on this recent work and use selected paradata, in addition to other auxiliary variables available for both respondents and nonrespondents, to estimate response propensities for active cases at subsequent contacts. Consider the following ingredients for computing a propensity-adjusted indicator of interviewer performance:

a) For each active case, assume that we know the probability that the next *contact* will generate a positive outcome; this probability is stored in the paradata record of the case, and is called p_{jc} (the probability of a successful outcome for case j at contact c).
b) The interviewer assigned the case contacts case j at the next call; if the outcome of the contact c is positive (a completed interview), then interviewer i is assigned a "deviation" score of $d_{ijc} = (1 - p_{jc})$; if the outcome if negative (a refusal, or a scheduled follow-up call or appointment[1]), then interviewer i is assigned the score $d_{ijc} = (0 - p_{jc})$; considering only a focus on contacts (ignoring

[1] Although appointments are certainly considered positive outcomes, they do not equate to completed interviews. Scheduled appointments are considered part of the paradata for a case's prior contact history, and are extremely important predictors of completed interviews in response propensity models (see Table 1), as would be expected. Cases with previously scheduled appointments are considered "easier" cases, because they have a high probability of yielding a completed interview at the next contact, and interviewers failing to secure an interview at the next contact would be penalized when using the PAIP method.

what cases are involved), we re-label this deviation score for a given contact *c* made by interviewer *i* as d_{ic}.

c) Interviewer *i* makes a total of C_i contacts during the data collection period; these deviation scores are then averaged over all contacts for each interviewer; the result for interviewer *i* is $\bar{d}_i = \frac{\sum_{c=1}^{C_i} d_{ic}}{C_i}$, or a propensity-adjusted interviewer performance (PAIP) score based on *contacts only*.

This performance indicator gives large credit to the interviewer who obtains success on very difficult cases and small penalty with failure on such cases. The performance indicator also gives a small credit to success on very easy cases and a large penalty to failure on such cases. Interviewers who happen to be assigned more difficult cases are rewarded for their success on high burden workloads with higher PAIP scores. In contrast, interviewers who happen to be assigned easier cases are less rewarded for their success on their low burden workloads. We note that response probabilities at subsequent contacts from part a) above are never known in practice, and need to be estimated using paradata and other auxiliary variables that are available for both respondents and nonrespondents. Response propensity models using strong predictors of cooperation with the survey request at the next contact are therefore essential to this method, and we evaluate the ability of a variety of paradata to predict response propensity in this paper.

We present two illustrations of the PAIP score: one using an area probability face-to-face sample survey with interviewers resident in primary sampling areas, and the other using several years of a monthly RDD telephone survey conducted in a centralized CATI facility. The analysis explores the link between cooperation and estimated difficulty of interviewing a sample case. The paper also compares the empirical estimates of PAIP scores to cooperation rates of interviewers, in an effort to evaluate whether the PAIP score adjusts in desirable ways for the differential difficulty of cases being assigned to different interviewers.

3. Two Illustrations of PAIP Scores and an Evaluation Relative to Interviewer-Level Cooperation Rates

3.1 Cycle 7 of the National Survey of Family Growth (NSFG)

The NSFG (Cycle 7 in the history of the National Fertility Surveys in the United States) is an ongoing area probability sample survey of US household members 15-44 years of age, with a 60-80 minute interview about sexual and fertility experiences, partnering, and family formation events (see http://www.cdc.gov/nchs/nsfg.htm). Every 12 weeks, or 84 days, a new sample of addresses is released for data collection in a national sample of 33 primary areas. The interview protocol is divided into two steps: a screening interview to determine whether there are one or more age-eligible (i.e., age 15-44 years) persons present in the selected household, and a main interview with an age-eligible respondent selected at random from all age-eligible respondents within the household. The 84-day period, or quarter, is divided into Phase 1 (first 70 days) and Phase 2 (last 14 days), where Phase 2 is a continuing data collection effort on a probability subsample of cases that have not been interviewed in Phase 1. In Phase 1 all sample persons are given $40 in cash upon agreeing with the interview; in Phase 2, all those persons are mailed $40 and then offered an additional $40 if they complete the main interview. At the end of a twelve-month period, 25 primary areas are rotated out of the sample and 25 new areas are rotated into the sample. Using the AAPOR RR4 calculation (AAPOR, 2008), with individual outcomes adjusted by base sampling weights, Cycle 7 of the NSFG achieved overall quarterly weighted response rates ranging from 70% to 80% across demographic subgroups in the first 10 quarters of the cycle. This paper examines the performance of 91 interviewers working in 83 different primary areas over the first 10 quarters of the survey (June 2006 to December 2008). Given that most NSFG interviewers work in only a single PSU,

this study attempted to include several PSU-level predictor variables that might explain variance in case difficulty among interviewers.

The PAIP scores examined for this survey reflect only the visits in which contact was established with someone at the sample housing unit, after completion of a screening interview. PAIP scores could also be computed for screening interviews separately, but we focus on contacts with cases having completed screening interviews only. Given the objective of the PAIP score (i.e., to incorporate propensity of obtaining an interview at the next contact into the interviewer performance measures), we chose to contrast the contacts that yielded a successful main interview, after respondent selection from the screening interview, with all other contact outcomes. The proportion of contacts that yield a main interview is approximately 0.32, meaning that about 32% of all contacts with cases selected for the main interview actually yielded a main interview. We note that this percentage differs from the overall response rate for a quarter, which (in general) describes the proportion of cases selected from the screening interview that were successfully interviewed. We also note that the PAIP score ignores differential interviewer performance in contacting cases; interviewer ability to obtain screening interviews, given initial contact; and various other aspects of the interviewer job. Clearly, the PAIP score could be altered to suit a wider focus, but the present score represents only one aspect of interviewer performance, and should be considered along with additional metrics to evaluate overall performance (Steinkamp, 1964).

Mated to the PAIP score is the propensity model, predicting for each contact the probability that a successful main interview will be conducted. The NSFG operates using a responsive design framework (Groves and Heeringa, 2006), and accordingly collects a very rich set of paradata on sample segments within primary areas (e.g., interviewer safety concerns), housing units (e.g., physical access impediments), calls (e.g., number of calls attempted), contacts (e.g., respondents asking questions during the contact), and interviewer observations (e.g., evidence of presence of young children) to enhance design decisions during data collection (Groves et al., 2009, pages 23-24). Given that the majority of these paradata are collected by interviewers during listing and screening operations, measurement error in the paradata is a real possibility, and research is currently ongoing to evaluate the quality of the paradata (e.g., West, 2010). For the purposes of this study, we use these paradata as predictors in our response propensity models, and acknowledge that measurement error in the paradata may attenuate some of the relationships of these predictors with response propensity (Stefanski and Carroll, 1985). We also acknowledge that not all survey programs will have a similar set of paradata to build response propensity models, although the collection of paradata has been shown to improve design efficiency (Groves and Heeringa, 2006).

Given our data set of calls where contact was established with a sample case selected from the screening interview (42,200 contacts in total from the first 10 NSFG quarters), we fit discrete-time hazard models to predict the probability of a successful interview at the next contact. In these analyses, the dependent variable was equal to 1 for contacts with successful interviews and 0 for contacts with other outcomes, a variety of paradata collected on the sample cases were used as predictors, and sample cases had as many records in the data set as contacts were made until a main interview was completed or the data collection was finished for the quarter. Table 1 presents the estimated coefficients of this discrete-time hazard model predicting the conditional probability that the next contact will produce a main interview (for statistically significant predictors only). Given the data structure described above, this model can be easily estimated with most standard logistic regression software (SAS/STAT PROC LOGISTIC Version 9.2 was used in the present study), as described by Allison (2010).

Table 1: Estimated Discrete-Time Hazard Model Coefficients and Standard Errors Predicting the Likelihood of Completing a Main Interview on the next visit or call attempt: NSFG Main Study, Cycle 7 (n = 42,200 contacts), and SCA Call History, 2003-2006

	NSFG Model	**SCA Model**

Predictor Variables	Estimate	SE	Estimate	SE
Intercept	-1.493**	0.049	-2.387**	0.638
Paradata: Calling History				
Number of Contacts with Resistance			-0.397**	0.021
Number of Contacts with no Resistance			-0.062**	0.006
Number of Non-Contacts			0.011**	0.001
Prior Contact Established	0.417**	0.050	0.103**	0.031
Sample Person Statements at Last Contact	-0.286**	0.085		
Number of Prior Calls	-0.044**	0.004		
Last Contact Produced Other Result	0.178**	0.033		
Last Contact Produced Soft Appointment	1.170**	0.072	0.585**	0.037
Last Contact Produced Hard Appointment	2.593**	0.039	1.338**	0.049
Last Contact Had Max Resistance	-0.458*	0.213	-0.428**	0.055
Days Since Last Contact (SCA Only)			-0.048**	0.004
Days Left in Month (SCA Only)			-0.006**	0.001
Interviewer Observations				
Spanish-language Screener	-0.168**	0.051		
Single-Person household	0.335**	0.043		
Evidence of Non-English Speakers	0.284**	0.094		
Evidence of Spanish Speakers	-0.330**	0.096		
Neighborhood Safety Concerns	0.086**	0.030		
Access Impediments	-0.118**	0.040		
Person Likely in Sexual Relationship	-0.114**	0.033		
Multi-unit Structure	0.071*	0.030		
Teenage Sample Person	0.086*	0.036		
County-level Census Variables (SCA Only)				
% Graduated from College			0.012**	0.003
% with Race = White			0.007**	0.001
Median Home Value in Thousands			-0.001**	<0.001
Northeast Census Region (vs. West)			-0.172**	0.041
Midwest Census Region (vs. West)			0.097**	0.037
South Census Region (vs. West)			-0.146**	0.037
Nielsen County A – Top 21 Metro Areas (vs. D)			-0.134*	0.053
Number of Contacts	42,200		84,858	
Nagelkerke Pseudo-R^2	0.332		0.074	
Percent Concordant Pairs	76.9%		65.3%	
Area Under the Curve (AUC)	0.773		0.657	

*p < 0.05; **p < 0.01

The model fits the data fairly well, given a set of paradata that were introduced into the NSFG design based on predictors of response propensity found in prior NSFG cycles (Lepkowski et al., 2006, pages 29-33); the pseudo-R^2 for the model is 0.33. The most powerful predictors reflect the prior calling experience, with higher interview propensities for contacts following hard appointments, contacts with fewer prior calls, and with many prior contacts. These findings support previous work by Beaumont (2005) examining the utility of using such paradata for repairing nonresponse bias. In addition, several interviewer observations were found to be strong predictors of response propensity. Higher response propensities were found in single-person units, English-speaking households, neighborhoods with evidence of languages other than Spanish and English spoken, neighborhoods with interviewer safety concerns, households without access impediments, and households where selected respondents were judged to not be in a sexually active relationship. Interestingly, urban areas and rural areas did not differ significantly in response propensities when controlling for the other paradata, suggesting that previous differences in response propensity between these areas reported in the literature may actually be driven by

variables similar to the paradata being collected in the NSFG (e.g., access impediments and safety concerns). We remind readers that these analyses were based on all contacts with sampled cases only, and that predictors of response propensity may vary in other analyses with different case bases.

From this model, for each contact, we could estimate the probability of obtaining an interview, p_{ijc}, by $\hat{p}_{ijc}$. This predicted probability was computed by setting the values of all predictors to the appropriate level, given the data record, and then transforming the estimated logit based on the fitted model (Table 1) to a probability. This approach therefore allows estimates of individual response propensities to change as a function of contact history (incorporating time-varying indicators of difficulty), and yields the estimated deviation score $\hat{d}_{ic}$ for the given contact *c* by interviewer *i* (which again reflects the difference between a binary indicator of whether the contact outcome was a main interview and the predicted response propensity). Given these scores for each contact, the overall PAIP score, representing the mean of all of the deviation scores from the contacts established by a given interviewer, can be computed for each interviewer.

As an evaluation of the propensity model and PAIP scores in general, Table 2 sorts all of the contacts in the NSFG data set by quintiles of the estimated probabilities of obtaining a main interview. Note that all contacts of all cases are included in the table; thus, if a sample address had three contacts, it appears three times in the table, with each of the three lying in the quintile appropriate to the propensity of that specific contact yielding a main interview. There are 42,200 contacts in the data set; those in the lowest 20% of estimated probabilities of a main interview, indeed, have only 13.3% yielding interviews. The percentage of contacts producing interviews steadily accelerates in a non-linear fashion across the quintiles, reaching a high of 78.0% for the highest quintile. Thus, as expected, the propensity model does a good job of discriminating contacts that are difficult for interviewers from those that are easy. Further, these results show how difficult some contacts can be for interviewers, and how easy other contacts can be.

Table 2: Percentages of NSFG Contacts Resulting in a Completed Interview and Standard Deviations of Cooperation Rates across Interviewers, by Quintile of Predicted Response Propensity

Quintile of Contact's Predicted Propensity of an Interview	Percentage of Contacts Resulting in an Interview (n = 42,200 contacts)	Standard Deviation of Cooperation Rates Among Interviewers for Final Contacts within the Quintile (number of interviewers)	Number of Contacts Falling into Quintile (Number of Final Contacts)
1 (Low)	13.3%	0.291 (n = 88)	8,439 (2,095)
2	17.5%	0.191 (n = 86)	8,441 (1,658)
3	19.5%	0.202 (n = 85)	8,435 (1,863)
4	30.8%	0.131 (n = 88)	8,445 (2,801)
5 (High)	78.0%	0.013 (n = 86)	8,440 (6,626)

It is common for surveys to compute the cooperation rate at the interviewer level, by the ratio of interviewed cases to all cases assigned to them that were contacted and eligible (see AAPOR, 2008). (If interviewers share sample cases, the interviewer obtaining the final result is often the "interviewer of record.") In Table 2 we compute the quintile-specific cooperation rate for each interviewer based on the results of the *final* contacts with assigned sample cases. For example, if an interviewer completed work on 20 cases and there were final contacts with four cases in each of the quintiles for that interviewer (based on their predicted probabilities of an interview at next contact), the interviewer would have five

cooperation rates computed, each based on the results of the final contacts with the four cases in each quintile.

The third column of Table 2 shows the standard deviation across interviewers of such cooperation rates for each quintile, and the fourth column of Table 2 shows the number of contacts falling into each quintile (in addition to the number of final contacts falling into each quintile). We expected that final contacts having a very high propensity of an interview would be successfully handled by almost all interviewers; that is, we would see lower standard deviations across interviewers in cooperation rates in the highest propensity quintile. Conversely, more difficult cases would lead to greater variability in cooperation rates across interviewers. Table 2 shows a relatively low standard deviation across interviewers for the highest propensity cases (standard deviation of interviewer cooperation rates = 0.013) and higher standard deviations for lower propensity quintiles, with a maximum standard deviation of 0.291 for the lowest propensity quintile. In short, easy cases (based on the estimated propensity model in Table 1) generated uniformly high response rates across all interviewers, and difficult cases produced large interviewer variation in cooperation rates. This is further support for the PAIP score approach as an evaluative tool of interviewers, and motivates the collection of additional paradata and auxiliary variables to improve the fits of response propensity models used to compute the PAIP scores.

Given the results above, we would expect the cooperation rates computed for each interviewer (based on final contacts) to be positively correlated with the mean predicted response propensity of cases assigned to the interviewer (also based on final contacts). That is, the more traditional way of evaluating interviewers based on final cooperation rates favors those interviewers who are assigned easier cases. We estimated the correlation between the interviewer-level cooperation rate and the mean predicted response propensity of cases assigned to the interviewer. The estimated correlation was $r = 0.47$ ($p < 0.001$), showing that interviewers with easier assignments do indeed have higher cooperation rates. When we estimate the correlation of the mean predicted response propensity for an interviewer with the PAIP score computed for the interviewer, the estimated correlation is $r = -0.51$ ($p < 0.001$), suggesting a very different story: interviewers working more difficult cases tend to have higher PAIP scores (meaning that success with difficult cases is being rewarded), while interviewers working easy cases tend to have lower PAIP scores (meaning that failures with easy cases are being penalized). The interviewer-level cooperation rate and the PAIP score are positively correlated but not strongly ($r = 0.34$), suggesting that the PAIP scores and the cooperation rates only have about 11% of their variance in common. As an alternative measure of association between these two performance measures, interviewers were grouped into quintiles on each measure, and the two quintiles were cross-tabulated. Only 38.5% of the interviewers fell into the same quintile on each measure (Simple Kappa = 0.23, 95% CI = 0.11, 0.36). In other words, the NSFG PAIP scores are capturing unique aspects of interviewer performance that are not being captured by the cooperation rates.

To summarize the results of these analyses of real NSFG data, we found several strong predictors of response propensity in a discrete-time hazard model being fitted to contacts only, using a rich set of paradata and interviewer observations collected in the NSFG as predictors. When using these predictors to predict the probability of a successful completed main interview at next contact, we can compute PAIP scores for each interviewer based on deviations between the actual outcomes from the contacts and the predicted probabilities of a successful interview, incorporating the difficulty of the assigned cases into the PAIP measures. We find that the percentage of contacts resulting in completed interviews increases substantially across percentiles of predicted response propensities, while the standard deviation of cooperation rates among interviewers tends to be highest for cases with low response propensities and lowest for cases with high response propensities. Estimates of correlations between PAIP scores, cooperation rates for interviewers, and mean predicted response propensities for cases assigned to interviewers provide evidence that the PAIP score is in fact capturing a unique aspect of interviewer performance that is not being captured by the traditional cooperation rate metric that fails to incorporate

case difficulty. In the next section, we consider similar analyses using data from a telephone survey to further evaluate the effectiveness of the PAIP score.

3.2 The Monthly Survey of Consumer Attitudes (SCA)

The Survey of Consumer Attitudes (SCA) is a monthly RDD household telephone survey interviewing randomly-selected adults about their attitudes and sentiments toward the economy and their personal financial condition (see http://www.sca.isr.umich.edu/main.php). Each month, about 300 new interviews are taken (in a separate part of the design, second-wave interviews are taken with those first interviewed six months earlier). The SCA analyses in this section are based on all call records resulting in contact from 2003 to 2006, using only the first-wave RDD sample cases. The response rate for the first-wave cases ranges from 43.2% to 47.8% (AAPOR RR2) during this period. As with the vast majority of centralized telephone surveys, interviewers working on the survey share the sample cases, with calling directed by CATI sample administration software. In contrast to the use of in-person interviewing to collect data from an area probability sample in the NSFG, the RDD sample design and the sharing of cases is expected to result in a more equitable distribution of difficult cases across the interviewers.

We replicate the analysis of the face-to-face NSFG data with the telephone SCA data. A much smaller set of paradata and auxiliary variables are generally available for predicting response propensity in telephone surveys, given the absence of interviewer observations on the neighborhoods and households and relatively sparse frame information. Nevertheless, we still consider a similar set of paradata collected in the SCA that reflects previous calling efforts (number of prior contacts, resistance levels on contacts, nature of prior appointments made, number of days since the last contact, and number of days left in the monthly data collection period). We also merged a detailed set of Census information into the data set containing SCA contact records. This was accomplished by first determining the Census tracts covered by each telephone number's exchange, and then linking aggregated Census tract-level data for the (possibly multiple) tracts covered by a telephone number's exchange into the contact data set (see Johnson et al., 2006 for a similar example). The merged Census variables in the SCA data set included aggregate measures (for all tracts covered by a given exchange) of household density, Nielsen county classification of the area, age distribution, race/ethnicity distribution, educational distribution, urbanicity, housing tenure distribution, household income distribution, percentage of phone numbers listed, and Census region. A discrete-time hazard model was fitted to all SCA contacts from 2003 to 2006 (84,858 contacts in total) using these paradata and the Census variables as predictors of a binary indicator for a successful completed interview given contact.

Table 1 presents estimates of the coefficients in the discrete time hazard model for the SCA contact data, alongside estimates for the NSFG contact data. Apparent in this table is the lack of available paradata for building this response propensity model relative to the NSFG, especially if Census variables had not been merged into the data set. Also of note is the fact that the paradata collected in the SCA tended to be much stronger predictors of interview completion than the Census variables, suggesting that increased focus on collection of paradata in phone surveys may be needed for improving the response propensity models. Only higher education, white ethnicity, lower median home values and Census region appear to be strong area-level predictors of response propensity, among all the Census variables considered. The ability of the SCA response propensity model to discriminate well among the difficulty of different contacts, not surprisingly, is much less than that of the NSFG model (pseudo-$R^2 = 0.074$).

Table 3 sorts all of the contacts in the four-year SCA data set by quintiles of the estimated probabilities of obtaining a main interview, based on the SCA response propensity model. Once again, all contacts of all cases are included in the table. In addition, we break results down by SCA year, to examine the stability of the results across the four years. There were 84,858 contacts overall in the data set across the four years. As with the NSFG above, quintile-specific cooperation rates were computed for each interviewer in each of the four years based on final contacts with the sampled telephone numbers. If a line's final call

fell into another quintile, then that line contributed to the interviewer's cooperation rate in the other quintile. As with the NSFG analysis, the percentage of contacts yielding an interview increased in a nearly monotonic fashion by propensity quintile (in each SCA year), suggesting that the model can predict the likelihood of an interview in a useful fashion despite the low pseudo-R^2 value.

Table 3: Percentages of SCA Contacts Resulting in a Completed Interview and Standard Deviations of Cooperation Rates across Interviewers, by Survey Year and Quintile of Predicted Response Propensity

Survey Year	Quintile of Contact's Predicted Propensity of an Interview	Percentage of Contacts Resulting in an Interview	Standard Deviation of Cooperation Rates Among Interviewers for Final Contacts within the Quintile (number of interviewers)	Number of Contacts Falling into Quintile (Number of Final Contacts)
2003 (21,228 contacts)	**1 (Low)**	12.2%	0.2422 (n = 57)	5793 (1832)
	2	11.6%	0.2572 (n = 58)	3924 (1064)
	3	15.5%	0.2659 (n = 54)	3840 (995)
	4	19.9%	0.2358 (n = 61)	3823 (1018)
	5 (High)	26.7%	0.2240 (n = 63)	3848 (1250)
2004 (21,232 contacts)	**1 (Low)**	7.6%	0.2064 (n = 46)	4015 (1477)
	2	11.4%	0.2082 (n = 49)	4217 (1207)
	3	16.1%	0.2207 (n = 52)	4466 (1231)
	4	20.4%	0.2147 (n = 50)	4506 (1374)
	5 (High)	29.3%	0.2308 (n = 50)	4028 (1521)
2005 (23,316 contacts)	**1 (Low)**	5.5%	0.1463 (n = 35)	5463 (1835)
	2	10.6%	0.2657 (n = 42)	4262 (1086)
	3	14.6%	0.2361 (n = 45)	4350 (1021)
	4	19.2%	0.2795 (n = 49)	4438 (1157)
	5 (High)	27.4%	0.2111 (n = 53)	4803 (1604)
2006 (23,072 contacts)	**1 (Low)**	6.6%	0.1948 (n = 36)	5690 (2003)
	2	10.8%	0.2787 (n = 42)	4569 (1222)
	3	15.9%	0.2160 (n = 39)	4315 (1094)
	4	18.1%	0.2005 (n = 39)	4205 (1098)
	5 (High)	29.0%	0.2050 (n = 45)	4293 (1519)

The third column of Table 3 shows the standard deviations across interviewers of the cooperation rates for cases with final contacts in each propensity quintile. Across the four SCA years, the relationship between the quintile and the standard deviation of cooperation rates is consistently much weaker than seen in Table 2 for the NSFG. There is relatively stable between-interviewer variation in cooperation rates across propensity quintiles, and there could be several reasons for this finding. First, in centralized telephone facilities, interviewers are given a more uniform mix of sample cases. This also means that individual interviewer cooperation rates are more uniform than in area probability surveys, where cooperation rates vary because of differences across primary areas. Second, unlike the NSFG, a much smaller proportion of final contacts in the SCA were with cases in the highest estimated propensity quintile, and a much higher proportion of final contacts were with cases in the lowest quintile. This would have the effect of reducing the standard deviation in the lowest quintile and increasing the standard deviation in the highest quintile (relative to the NSFG), assuming that sums of squared deviations from the mean in each quintile were similar to those found in NSFG.

The mean predicted response propensities for each interviewer's final contacts were also computed (again, to indicate difficulty of cases assigned). The mean response propensities based on the propensity model fitted to the contacts and the cooperation rates for the interviewers in each year were moderately correlated (with correlations running between 0.42 and 0.60 across the four years), suggesting that cooperation rates tended to be similar to average response propensities (consistent with the NSFG results). That is, interviewers in centralized phone facilities who disproportionately call easy cases achieve higher cooperation rates. Correlations of the mean response propensities and the computed PAIP scores ranged from -0.09 to 0.27 across the four years, with only one correlation (0.27) being significant at $p < 0.05$. These negligible correlations suggest that PAIP scores do not provide a strong indication of response propensity in telephone surveys, but this could be a function of the poorer response propensity model relative to NSFG (where a strong negative correlation was found, indicating that the PAIP score tends to penalize interviewers with easier cases). Finally, the correlations between interviewer cooperation rates and the computed PAIP scores ranged between 0.39 and 0.63 across the four years. These moderate correlations once again suggest that the two indicators are measuring somewhat different phenomena, with at most 40% of their variance shared in common.

4. Discussion

Cooperation rates and response rates obtained by interviewers, based on their assigned cases, are commonly used to evaluate interviewer performance in surveys. Some sample cases are relatively easy to interview, and others are more difficult. If interviewers were assigned sample cases at random, using their cooperation rates to evaluate them would be a fair process. However, interviewers given more difficult cases than others are disadvantaged by such a process. Using propensity models for the likelihood of obtaining an interview on a given contact, we showed the empirical magnitude of this disadvantage for two different surveys.

The use of the PAIP score proposed in this paper attempts to rectify this inequity by evaluating interviewers based on how their achievement of interviews exceeds what was expected for a given contact, given the attributes and prior calling outcomes for an assigned case. The PAIP score requires the survey researcher to estimate a response propensity model, which can be used to predict the probability of a contact yielding an interview on a given call. Although collecting measures on the variables used as predictors in these models may require additional effort, most survey organizations already collect information on similar auxiliary variables for use in nonresponse adjustments. Given an information system that can collect and update information on the auxiliary variables and contact outcomes on a daily basis, these models can be estimated each day from call record data (organized into a data set of contacts) with standard statistical software capable of fitting logistic regression models. Future research might also consider alternative methods of predicting response propensity, such as classification trees (which quickly

determine important interactions between input predictor variables). Any modeling technique that enables prediction of response propensities for active cases as a function of input predictor variables could be used to compute PAIP scores, but the sensitivity of PAIP scores to alternative models should be closely examined in future applications.

In this study, we found that a propensity model fitted to data collected in a personal interview household survey had a much better fit than a propensity model fitted to data collected in a telephone survey. The in-person survey format enables the collection of a much richer set of paradata, including interviewer observations, and this was likely the reason for the large differences in fit between the two models. If interviewer observations are shown to be significant predictors of response propensity in other applications of the PAIP score, survey organizations should be wary of interviewers noticing that their observations influence the predicted difficulty of active cases. As a result, interviewers may attempt to continuously enter observations suggesting that cases are more difficult than they are to avoid penalties for unsuccessful contact attempts. Having a system in place that provides interviewers (or possibly a random sample of interviewers) with continuous feedback on the accuracy of their observations would help to minimize this problem. More generally, additional research into the quality and validity of paradata collected in in-person surveys is certainly needed, as substantial errors in paradata may reduce the performance of the response propensity models used to compute PAIP scores. Initial work is ongoing in this area (West, 2010).

Linked Census information was not found to substantially improve the fit of the response propensity model fitted to the data from the telephone survey, which also collected paradata on previous calling efforts similar to the paradata collected for the in-person survey. We acknowledge that the PAIP score relies on a strong propensity model that can accurately predict the probability of a successful interview at next contact, and the findings in this paper emphasize the need for research into new methods for collecting paradata in telephone surveys to improve response propensity models. Despite the poor fit of the propensity model in the telephone survey, this paper has demonstrated that purifying interviewer-level cooperation rates with PAIP scores allows the survey manager to make management interventions with greater clarity that account for discrepancies in the difficulty of cases assigned to interviewers. Improving the fits of response propensity models and the collection of paradata that can predict response propensity in telephone surveys should receive increased research attention so that telephone survey managers can make more informed decisions about interviewer performance. This becomes especially important in telephone surveys (and also face-to-face surveys) where many cases are finalized on the very first call (e.g., a completed interview or a hard refusal), and extensive paradata on prior calling efforts are not available. Listed status of a telephone number might be helpful here.

This paper merely introduced the notion of a PAIP score and provided two examples of its use. We acknowledge that the PAIP score would only be one metric out of many that might be used to provide a full picture of interviewer performance (Steinkamp, 1964), but we feel that the PAIP score is unique in terms of its potential to adjust for the difficulty of cases assigned to an interviewer. There are many other developments regarding PAIP scores that are meritorious. First, especially for face-to-face area probability sample surveys, a PAIP score could be developed to evaluate interviewers based on their ability to contact sample cases. Second, we used a deviation as the base of the PAIP score: $d_{ijc} = (1 - p_{jc})$

or $d_{ijc} = (0 - p_{jc})$. Looking at ratios rather than differences might be useful. Third, further analyses may

discover interviewers who are essentially outliers. One outlier condition is interviewers who succeed especially well on low propensity contacts. These interviewers might be used as special resources for the data collection, especially in two-phase designs dedicating additional resources toward securing cooperation from active nonrespondents in the second phase. Fourth, in ongoing survey operations,

attempting interventions that reduce the variation in PAIP scores across interviewers would be a useful management focus. Fifth, survey researchers attracted to PAIP scores have yet another reason to take the development of paradata seriously, as they can be useful predictor variables in the response propensity models used to compute PAIP scores. Finally, future studies of surveys with record data available for respondents could examine the associations of interviewer-level PAIP scores with 1) mean survey responses for interviewers and 2) mean response deviations for interviewers. These types of analyses would indicate whether interviewers with different PAIP scores tend to collect different distributions of survey responses, and whether PAIP scores could also serve as indicators of data quality.

References

Abbott, C.L., Yost, B.A., and Harding, J.L. (2004). Measures of Personality Type and Interviewer Performance: Tools for Interviewer Training. *Paper presented at the 2004 Annual Meeting of the American Association for Public Opinion Research, Phoenix, AZ, 5/16/2004.*

Allison, Paul (2010). *Survival Analysis using the SAS System, Second Edition*. SAS Publishing, Cary, NC.

American Association for Public Opinion Research (2008), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 5th edition*. Lenexa, Kansas: AAPOR.

Anton, J. (1997), *Call Center Management*, West Lafayette, IN: Purdue University Press.

Barioux, M. (1952). A method for the selection, training, and evaluation of interviewers. *Public Opinion Quarterly*, 16, 128-130.

Beaumont, J-F. (2005). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, 31(2), 227-231.

Berk, M.L. and Bernstein, A.B. (1988). Interviewer Characteristics and Performance on a Complex Health Survey. *Social Science Research*, 17, 239-251.

Biemer, Paul, Herget, Deborah, Morton, Jeremy, and Gordon Willis. (2000). The feasibility of monitoring field interview performance using computer audio recorded interviewing (CARI). *Proceedings of the Survey Research Methods Section of ASA*, 2000 Joint Statistical Meetings, 1068-1073.

Cannell, C.F., Lawson, S.A., and Hausser, D.L. (1975). *A Technique for Evaluating Interviewer Performance: A Manual for Coding and Analyzing Interviewer Behavior from Tape Recordings of Household Interviews*. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan.

Chun, Y., Kwanisai, M., and Hamilton, L. (2010). Paradata for Response Propensity Models and Nonresponse Adjustment in the National Assessment of Educational Progress: Social Isolation Theory as a Modeling Navigator. *Proceedings of the Survey Research Methods Section of the American Statistical Association, Joint Statistical Meetings*, Vancouver, British Columbia, August 2010.

Couper, M.P. (1998), "Measuring Survey Quality in a CASIC Environment." Invited paper presented at the Joint Statistical Meetings of the American Statistical Association, Dallas, August.

Couper, M.P. and Lyberg, L. (2005). The use of paradata in survey research. *Proceedings of the 55th Session of the International Statistical Institute*.

Curtin, R., Presser, S. and Singer, E. (2005). Changes in telephone survey nonresponse over the past Quarter century. *Publ. Opin. Q.*, 69, 87–98.

de Leeuw, Edith, and de Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. Chapter 3 in Groves, R.M. et al., *Survey Nonresponse*. Wiley.

Durand, Claire. (2008). Assessing the usefulness of a new measure of interviewer performance in telephone surveys. *Public Opinion Quarterly*, 72(4), 741-752.

Durand, Claire. (2005). Measuring interviewer performance in telephone surveys. *Quality and Quantity*, 39(6), 763-778.

Edwards, Sandra, Slattery, Martha L., Mori, Motomi, Berry, T. Dennis, Caan, Bette J., Palmer, P. and John D. Potter. (1994). Objective system for interviewer performance evaluation for use in epidemiologic studies. *American Journal of Epidemiology*, 140(11), 1020-1028.

Fowler, Floyd, Jr. (2008), *Survey Research Methods*, Fourth Edition, Thousand Oaks, CA: Sage Publications.

Groves, Robert, and Couper, Mick (1998), *Nonresponse in Household Interview Surveys*, New York: Wiley.

Groves, R.M., and Heeringa, S.G. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs. *J.R. Statist. Soc. A*, 169, Part 3, 439-457.

Groves, R.M., Mosher, W.D., Lepkowski, J. and Kirgis, N.G. (2009). Planning and development of the continuous National Survey of Family Growth. National Center for Health Care Statistics. *Vital Health Statistics*, 1(48).

Herget, Deborah, Biemer, Paul, Morton, Jeremy and Kelly Sand. (2001). Computer Audio Recorded Interviewing (CARI): Additional Feasibility Efforts of Monitoring Field Interview Performance. *Proceedings of the 2001 Federal Committee on Statistical Methodology, Session III-A.*

Hox, Joop J. and de Leeuw, Edith D. (2002). The influence of interviewers' attitude and behavior on household survey nonresponse: an international comparison. Pp. 103-120 in R.M. Groves, D.A. Dillman, J.L. Eltinge & R.J.A. Little (Eds.) *Survey Nonresponse*. New York: Wiley.

Hox, Joop J., de Leeuw, Edith D. and Kreft, Ita G.G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. In: P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds.) *Measurement Errors in Surveys*. New York: Wiley.

Jocelyn, W. and Baribeau, B. (2010). A Study of Modeling Longitudinal Nonresponse Using Paradata in the Survey of Labour and Income Dynamics (SLID). *Proceedings of the Survey Research Methods Section of the American Statistical Association, Joint Statistical Meetings*, Vancouver, British Columbia, August 2010.

Johnson, Timothy P., Young, IK Cho, Richard T. Campbell, and Allyson L. Holbrook. (2006). Using community-level correlates to evaluate nonresponse effects in a telephone survey. *Public Opinion Quarterly*, 70(5), 704-719.

Kennickell, Arthur B. (2006), "Who's asking? Interviewers, Their Incentives, and Data Quality in Field Surveys," Survey of Consumer Finances Working Paper, SCF Web Site: http://www.federalreserve.gov/pubs/oss/oss2/scfindex.html.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M., and Raghunathan, T.E. (2010). Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Nonresponse: Examples from Multiple Surveys. *Journal of the Royal Statistical Society - Series A*, 173, Part 3, 1-21.

Laflamme, Francois and St. Jean, Helene. (2011). Proposed indicators to assess interviewer performance in CATI surveys. *Proceedings of the Survey Research Methods of the American Statistical Association, Joint Statistical Meetings*, Miami, Florida, August 2011.

Lepkowski, J.M., Mosher, W.D., Davis, K.E. et al. (2006). National Survey of Family Growth, Cycle 6: Sample design, weighting, imputation, and variance estimation. National Center for Health Statistics, Vital Health Statistics, 2(142), July 2006.

Manheimer, Dean and Hyman, Herbert. (1949). Interviewer performance in area sampling. *Public Opinion Quarterly*, 13(1), 83-92.

Morton-Williams, Jean. (1993). *Interviewer Approaches*. Aldershot: Dartmouth Publishing Company Limited.

Olson, K. (2006). Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias. *Public Opinion Quarterly*, 70(5), 737-758.

O'Muircheartaigh, C. and Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society, Series A*, 162, Part 3, 437-446.

Peng, David, and Karl Feld. (2011). Quality Control in Survey Research Today. *Survey Practice*, April: www.surveypractice.org.

Sheatsley, P.B. (1949). The influence of sub-questions on interviewer performance. *Public Opinion Quarterly*, 13(2), 310-313.

Singer, E., Frankel, M.R., and Glassman, M.B. (1983). The Effect of Interviewer Characteristics and Expectations on Response. *Public Opinion Quarterly*, 47, 68-83.

Snijkers, Ger, Hox, Joop J. and De Leeuw, Edith D. (1999). Interviewers' tactics for fighting survey nonresponse. *Journal of Official Statistics*, 15, 185-198.

Stefanski, L.A., and Carroll, R.J. (1985). Covariate Measurement Error in Logistic Regression. *The Annals of Statistics*, 13(4), 1335-1351.

Steinkamp, S. (1964). The identification of effective interviewers. *Journal of the American Statistical Association*, 59, 1165-1174.

Stoop, I., Billiet, J., Koch, A. and Fitzgerald, R. (2010). *Improving Survey Response: Lessons Learned from the European Social Survey*. Wiley.

Sudman, S. (1966). Quantifying Interviewer Quality. *Public Opinion Quarterly*, 30(4), 664-667.

Summers, G.F. and Beck, E.M. (1973). Social Status and Personality Factors in Predicting Interviewer Performance. *Sociological Methods and Research*, 2(1), 111-122.

Tarnai, J, and Moore, D.L. (2008). Measuring and Improving Telephone Interviewer Performance and Productivity. Chapter 17 in *Advances in Telephone Survey Methodology*, Editors Lepkowski, J.M., Tucker, C., Brick, J.M., de Leeuw, E.D., Japec, L., Lavrakas, P.J., Link, M.W., and Sangster, R.L. John Wiley & Sons, Hoboken, New Jersey.

Waite, A. (2002), *A Practical Guide to Call Center Technology*, New York: CMP Books.

West, B.T. (2010). An Examination of the Quality and Utility of Interviewer Estimates of Household Characteristics in the National Survey of Family Growth. *Paper presented at the 2010 Annual Meeting of the American Association for Public Opinion Research, Chicago, IL, 5/14/2010.*

Wiggins, Richard D., Longford, Nicholas T., and O'Muircheartaigh, Colm A. (1992). A variance components approach to interviewer effects. Pp. 243-254 in *Survey and Statistical Computing*. A. Westlake, R. Banks, C. Payne & T. Orchard (Eds). Amsterdam: North-Holland.