Paradata, Interviewing Quality, and Interviewer Effects

by

Sharan N. Sharma

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey Methodology)
in The University of Michigan
2019

Doctoral Committee:

Professor Michael Elliott, Chair
Professor Mick Couper
Professor Frauke Kreuter
Assistant Research Scientist Zeina Mneimneh
Associate Professor Brady West

Sharan Sharma
snsharma@umich.edu
ORCID iD: 0000-0003-3434-5181

To Sharada

To all my teachers

To Archana

To survey interviewers

Acknowledgements

I consider myself extremely fortunate to have had Dr. Michael Elliott as my advisor. Mike was one of the instructors in the Ph.D. seminar where the idea for this dissertation was first proposed. Not only did he encourage me to listen to my intuitions and take the idea forward, he even made time for a special reading course on the topic. Mike's approach of giving me the freedom to do what I wanted, but also gently guiding me in the right direction, helped me immensely. I thoroughly enjoyed the 100+ meetings with him when we discussed statistics, methodology, and life; I suspect I will have withdrawal symptoms on Friday afternoons for quite a while after my graduation.

Dr. Frauke Kreuter was my other Ph.D. seminar instructor. I am always impressed with the way her advice goes to the heart of an issue. Her advice has helped me learn not only how to do research but also how to position my work. At some point, I also want to learn from her how to wear many hats and not crumble under their weight.

It was a privilege to have Dr. Mick Couper, the person who invented the term this dissertation deals with, on my committee. His insights, based on vast experience, were very helpful. Mick also served on my comprehensive exam committee; his recommendation to understand underlying mechanisms and not rush into the final analysis was directly responsible for Chapter 2 of this dissertation.

Dr. Brady West has been super supportive of all my work and his recommendations have opened a lot of opportunities. I'm grateful for the effort he puts into his teaching and always keeping student interests in mind. If I ever find myself taking a piece of writing lightly, I only have to visualize Brady to jerk me back into attention.

My first book chapter ever was co-authored with Dr. Zeina Mneimneh, who is probably the most patient co-author ever. Ever approachable and someone with great ideas, I could always count on her to give detailed and helpful comments on my work. Discussions with Zeina helped me structure the dissertation the way it appears now.

I'm thankful to Dr. Katherine McGonagle and the PSID team for giving me access to the data used for this dissertation. Patty Maher, Grant Benson, Lisa Holland, and Shonda

Kruger-Ndiaye from Survey Research Operations patiently dealt with my numerous questions. Shonda, in particular, spotted and helped correct a critical issue in the data that I was using thus helping me stay on track.

Dr. T.E. Raghunathan provided the impetus for my doctoral plunge when I was undecided; I look up to him for his deep knowledge and wisdom, and ability to live *Vedanta*. Dr. James Lepkowski and Dr. Steve Heeringa served on my comprehensive exam committee and gave excellent advice, and Dr. Fred Conrad was always encouraging and enthusiastic about my research. Beth-Ellen Pennell singularly mentored my international research interests, leading me to many significant opportunities. Jill Esau, Nancy Oeffner, Patsy Gregory, Jodi Holbrook, Elisabeth Schneider (who would send New Year cards without fail), and Sumi Rajendran, ably and in a most friendly way, provided all administrative help. I was fortunate to walk the journey along with have an excellent set of fellow-students. I specifically would like to thank Dr. Mengyao Hu for being a true friend and show me how to not overthink issues.

My deep gratitude goes to LV Krishnan, CEO of TAM India, for being my boss, friend, and guide for the last 18 years. I've never seen another person with the same combination of business skills and a grand sense of humanity.

My grandparents, parents (Vathsala Sharma and Narasimha Sharma), and sister (Shubha Sharma) have, among many other things, provided a loving family environment. What more can one ask for? My wife, Archana, has truly been my 'life support'; this dissertation belongs to her as much as to myself. Our children, Vishnu and Tara, have brought us so much joy; they constantly remind me of the 'bigger life'. My parents-in-law (M. Radhamani and P.K. Subramanya) are an inspiration in how to live an energetic life; I thank them for their blessings. At a time when I wasn't sure how to fund my studies, Captain H.S. Satyanarayana (my uncle) simply broke his bank deposits and offered me the money; gestures like these keep one motivated for a long time. Chetana and Shankar are a constant source of affection and warm wishes. Zarana and Karthik enthusiastically helped us settle us down in the beautiful town of Ann Arbor when we got to the U.S. and continue to help us be more 'street-smart'.

Finally, I thank Shri Ganesh Prasad for being the quintessential friend, philosopher, and guide, and taking me back to the 'good old days' in each of our long conversations.

I apologize to the many others I have missed mentioning but who have helped me successfully complete this major educational milestone.

Table of Contents

List of Tables

List of Figures

Abstract

A vast literature on interviewer effects (interviewer measurement error variance) is devoted to the estimation of these effects and understanding their causes and associated factors. However, consideration of interviewer effects in active quality control (QC) does not seem widespread, despite their known effect on reducing precision of survey estimates. We address this gap in this dissertation with the overarching goal of using item-level paradata (keystrokes and time stamps generated during the computer-assisted interviewing process) in a systematic manner to develop an active interviewer monitoring system in order to control interviewer effects. The dissertation is structured around exploring associations between paradata, indicators of interviewing quality, and interviewer effects. Our hypothesis is that different levels of interviewing quality cause different paradata patterns. Differing levels of interviewing quality also result in different between-interviewer response means even after controlling for respondent characteristics, leading to interviewer effects. Thus, interviewing quality is conceptualized as a common cause of both interviewer effects and paradata patterns, making it possible for us to think about paradata patterns as being potentially effective proxies of interviewer effects.

Little is known about what paradata say about the actual quality of an interview. This is explored in Chapter 2 where we use paradata patterns to either predict the proportion of flags in an interview (interview-level analysis) or the occurrence of a QC flag for an item (item-level analysis). The results show that paradata patterns have strong associations with interviewing quality. A key finding is that a multivariate approach to paradata use is necessary.

Chapter 3 turns to investigating associations of indicators of interviewing quality with interviewer effects. Survey quality control (QC) systems monitor interviewers for their compliance with interviewing protocol. But what is not very clear is if deviations from protocol are also associated with interviewer effects. While the results of our analysis show moderate associations in this regard, we find that when QC variables are complementary to other interviewer-level characteristic variables; when used together, a fair magnitude of interviewer variance can be explained.

Based on the foundations laid by Chapters 2 and 3, Chapter 4 uses paradata to directly predict interviewer effects. We find that paradata are fairly strong predictors of interviewer effects for the items we analyzed, explaining more than half the magnitude of interviewer effects on average. Also, paradata outperformed interviewer-level demographic and work-related variables in explaining interviewer effects. While most of the focus in the literature and practice has been on time-based paradata, e.g., item times, we find that non-time based paradata, e.g., frequency of item revisits, outperform the time-based paradata for a large majority of items. We discuss how survey organizations can use the dissertation findings in active quality control. All our analyses use data from the 2015 wave of the Panel Study of Income Dynamics.

Chapter 1

Introduction

1.1 Dissertation goal

The overarching goal of this dissertation is to examine if paradata can be used systematically to develop an active interviewer monitoring system in order to measure and manage interviewer effects. To achieve this goal, we explore associations between paradata, indicators of interviewing quality, and interviewer effects.

We use the term 'paradata' in the specific sense that it was first defined by Couper (1998), i.e., keystrokes and time-stamps that are generated in the course of a Computer-Assisted Interviewing (CAI) survey. By an 'active' interviewer monitoring system, we mean a system that enables a survey manager to quickly detect possible interviewing issues and undertake remedial actions before more errors are committed during fieldwork. By 'interviewer effects', we mean interviewer measurement error variance (except in Chapter 3 where we also use the term in its bias sense).

1.2 Motivation

A vast majority of professionally-run surveys rely on interviewers for data collection. In these surveys, interviewers play a vital role in contacting respondents, soliciting their cooperation, motivating them, and trying to ensure that accurate responses are obtained. However, interviewers are also a source of error. First, interviewers vary in their ability to obtain answers from respondents potentially giving rise to non-response error. Second, leaving aside behavior like falsification, the way a specific interviewer asks questions, probes respondents' answers, and gives feedback, induces responses that could be differ-

ent had another interviewer conducted the interview. This results in different expected response means across interviewers, even after factoring out differences in geographic and respondent profiles. These interviewer-induced measurement errors are called 'interviewer effects'. The associated intra-interviewer correlations (ρ_{int}; Kish 1962) are typically small in magnitude; Tucker (1983) finds the mean $\hat{\rho}_{int}$ computed across several variables for 11 telephonic surveys to be 0.004, and Groves (1989) finds values below 0.02 most common for 130 statistics computed from 10 personal interview surveys. But even these small values can substantially reduce the precision of an estimate since they increase the variance of descriptive estimates by a function of their product with interviewer workload.

One of the methods to control interviewer effects is the use of 'standardized interviewing' (Fowler and Mangione 1990) in which interviewers are trained to conduct all interviews in a standard manner so that error-inducing behavioral differences are reduced, if not eliminated. To track compliance with interviewing protocols, methods have been developed to monitor interviewers' behavior. These include accompanying interviewers on field (for face-to-face interviews), listening-in on interviews (for centralized telephonic interviews), re-interviews with respondents, or recording interviews and listening to them later. The last option is attractive since it is less obtrusive and facilitates the coding of interviewer behavior so that quantifiable information can be extracted.

However, these monitoring methods have two primary shortcomings. First, they do not *directly* link to estimates of interviewer effects; while behaviors are monitored through these methods, there is rarely any association made between the monitored behaviors and possible survey error. The second shortcoming is related to the efficiency of these methods. With the advent of Computer Audio-Recorded Interviewing (CARI; Biemer et al. 2000; Thissen et al. 2008; Mitchell et al. 2008; Thissen 2014), potentially all interviews can be recorded. But it still requires human coders to listen to these interviews and code them. Since it is expensive and time-consuming to listen to all recordings, survey managers typically sample a small proportion of cases to monitor, potentially missing truly problematic interviews. Moreover, trained interviewers rarely conduct entire interviews badly; different interviewers struggle with different types of items and respondents. Ideally, survey managers would be equipped with a method that links each item in an administered interview with a likelihood-of-measurement-error metric. This would guide the selection of which recording slice to listen to, resulting in a more efficient and effective (tailored) approach to giving feedback to interviewers. We aim to develop such a method using paradata.

1.3 Paradata and interviewer monitoring

Why would we expect paradata to be useful to monitor interviewers for measurement error? Consider two alternate interview scenarios that involve the same respondent but two different interviewers.

Scenario 1

INTERVIEWER1: Thinking back on the past week, on an average, for how many hours a day did you watch TV ?
RESPONDENT: Am not sure.
INTERVIEWER1: Okay. [registers a 'Don't know']

Scenario 2

INTERVIEWER2: Thinking back on the past week, on an average, for how many hours a day did you watch TV ?
RESPONDENT: Am not sure.
INTERVIEWER2: Maybe I could repeat the question to help you answer the question. [slows down pace] Thinking back on the past week, on an average, for how many hours a day did you watch TV ?
RESPONDENT: What do you mean by "on average"?
INTERVIEWER2: In the last seven days, there might have been days you could have watched less TV and some days more TV. But if I asked you to give me one number that stands for your daily TV viewing over the week, what would that be? Please take your time to recall your TV viewing last week and answer the question.
RESPONDENT: About two and a half hours.

The two interviews obtained very different responses. Interviewer1 does not take the effort to probe when faced with a non-response. On the other hand, interviewer2 undertakes the right steps by first repeating the question to the respondent, clarifying the question in a neutral fashion, and giving encouraging feedback. These actions are likely to have yielded an accurate response. Such interviewer behavior tends to be consistent as noted by Fowler and Mangione (1990, p.45): "Some interviewers obtain more answers than others on a consistent basis because they consistently probe for more answers, and that affects the data".

Broadly, a respondent goes through the process of comprehending the question (Comprehension), recalling the relevant information from memory (Retrieval), combining various recalled information in order to answer the question (Judgment), and finally communicating the answer (Tourangeau et al. 2000). An interviewer who is speeding through the

process could make comprehension difficult, and even if the respondent did comprehend the question, the rushed behavior could give cues that the respondent need not bother responding carefully (Fowler and Mangione 1990, p.71). On the other hand, careful probing would encourage better cognitive processing by the respondent resulting in better data quality. However, this effort would also tend to be associated with higher item times. In other words, item times (paradata in general) could be capturing interviewer behaviors that are associated with survey error.

These intuitions are not new. As early as 1964, Steinkamp looked at average interview length and variability in interview length to assess interviewer performance. Another example comes from Groves (1983) who said that data sets that count the number of back-ups during an interview "are badly needed [...] to measure the behaviors of interviewers during questioning", the implicit assumption being that more back-ups could mean more measurement error. Despite these long-standing intuitions, there seem to be no empirical links established between paradata and interviewing quality. Current quality control efforts continue to use methods based on intuitions such as focusing on interviews that are completed before a certain threshold minimum time (e.g., Cheung et al. 2016). While such heuristics can be useful they are inherently subjective and unlikely to be optimal. With technological advancement, the ease of obtaining paradata has increased. Ironically, this sometimes becomes an obstacle since survey managers are inundated with paradata and find it challenging to separate signal from noise. This requires research on the properties of paradata but, so far, as pointed out by Couper and Kreuter (2013), "relatively little attention has been paid to keystroke or item-level paradata" and "the absence of research on the large-scale use of measurement-error-related paradata in interview surveys is unfortunate".

We attempt to fill these gaps in interviewer monitoring and paradata research by analyzing associations between paradata, interviewing quality and interviewer effects. If such associations exist, the strengths of paradata, e.g., accuracy and being available for all respondents quickly at low cost, can be utilized to predict items which an interviewer may be struggling with. Quality control staff could then listen to these specific recordings and draw up a tailored training program which could be in the form of a simple low-cost exercise such as calling up interviewers and giving them feedback specific to their areas of weakness. The feedback can be followed-up by a re-evaluation after more fieldwork is completed.

1.4 Ideal survey and data

As a useful benchmarking exercise, we picture an ideal survey and dataset for our research. Given the above objectives, these would have the following characteristics:

- Paradata: The survey would be a CAI survey in which the instrument would capture and store every keystroke made, along with the associated date and time stamp. It is critical to have external software to be able to parse these paradata. Figure 1.1 reproduces an example from Cheung et al. (2016) which displays the raw paradata as captured by the CAI instrument; transforming such data for analysts' use is not trivial.

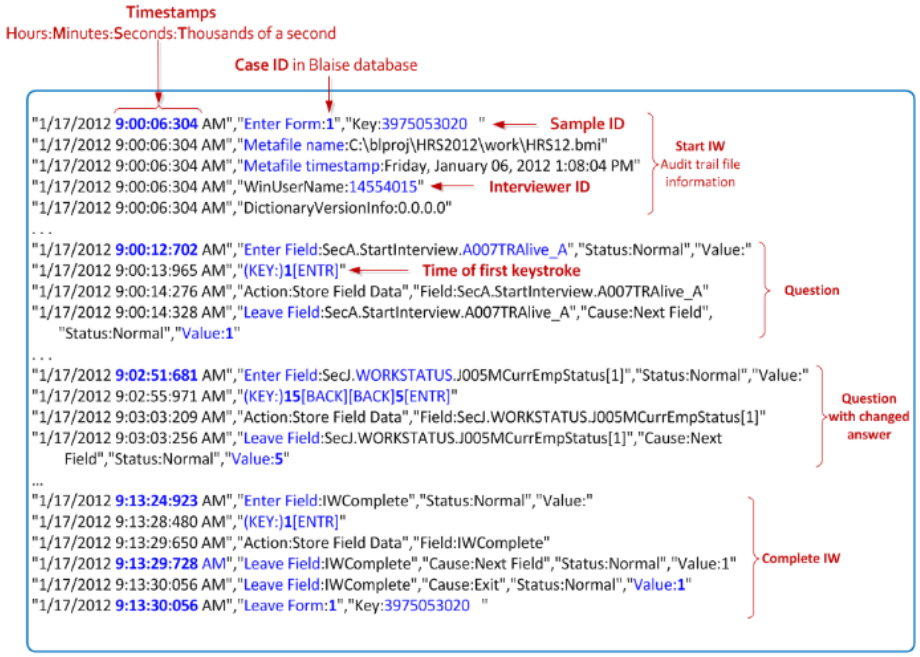

Figure 1.1: Example of raw paradata. Source: Cheung et al. 2016.

- Behavior codings: The survey would record and code interviewer behaviors for all its interviews.

- Sample Size: The survey would deploy a large number of interviewers, each with a large workload. Since our interest is in estimating interviewer measurement error variances, the number of interviewers is more important than the overall sample size (Hox 1998; Maas and Hox 2005, Raudenbush 2008, p.228-229). However, very small workloads (even if the number of interviewers is large) can create problems since likelihood ratio tests that rely on asymptotic results will not hold, interviewer effects are underestimated (Raudenbush 2008, p.225) but precision of the estimates is overestimated especially when interviewer effects are small (Raudenbush 2008, p.227), and numeric evaluation of integrals in the case of non-linear models would be

difficult (Raudenbush 2008, p.209,234). Biemer and Lyberg (2003, p.168) indicate a minimum 20/50 size (i.e., 20 interviewers with a minimum workload of 50 cases each), Hox (1998) suggests a 100/10 rule if there is a special interest in variance components, while Maas and Hox (2005) indicate that a sample size of 50 with workloads even as small as 5 is sufficient. Using the maximum of the indicated sizes in these suggestions, our ideal survey would have at least 100 interviewers with a workload of at least 50 interviews per interviewer (a total size of at least 5000 respondents).

- Interpenetrated design; the survey would employ an interpenetrated sample design (Mahalanobis 1946) to avoid the confounding of sampling and measurement errors.
- True values; true values for factual items would be available to compute both measurement error biases and measurement error variances.

1.5 Survey and data used for the dissertation

We searched for surveys that came close to the above characteristics. To the best of our knowledge, there is no survey that simultaneously meets the last two of the above characteristics. True values for factual items might be accessible for some federal surveys via special Research Data Center enclaves but access to these data would be very difficult, let alone access to paradata and interviewing quality data.

We requested for, and obtained, access to data from the Panel Study of Income Dynamics (PSID), a nationally representative survey of families and individuals in the U.S., conducted via Computer Assisted Telephone Interviewing (CATI); a small proportion of interviews (approximately 2.5%) are conducted face-to-face but we focused only on the CATI interviews for this dissertation. The survey is designed and executed by the Survey Research Center (SRC), Institute for Social Research (ISR), University of Michigan, Ann Arbor. We used the 2015 wave of the PSID where 9048 respondents were interviewed by 96 interviewers, thus meeting the above sample size requirements. While paradata are available for all interviews, interviewing quality data can be used only for a sample of 555 interviews. While not being a very small size, this reduced sample size (as compared to the full PSID sample) impacts the power to detect effects in analyses involving interviewing quality. PSID's telephonic mode is an advantage since there are no confounding area effects to contend with. Also the PSID has many questions on the economic situation of the household. Such questions can be sensitive and be susceptible to interviewer effects (Schaeffer et al. 2010; West and Blom 2016). On the other hand, our effects could also be dampened since the PSID is a panel survey and the majority of respondents would

be familiar with the questions. A Memorandum of Understanding (MoU) with SRC's Survey Research Operations governs access to data; while we have access to data on interviewer evaluations, the MoU does not allow access to the recordings themselves. This somewhat prevents us from drawing qualitative insights to supplement model results. Data on interviewers' sex, age, and education levels have also been provided but data on interviewers' experience levels are not accessible. Detailed descriptions of the survey and data are given in the relevant dissertation chapter sections.

1.6 Dissertation structure

This dissertation consists of 5 chapters including the current one. Our starting point is the hypothesis that different levels of interviewing quality cause different paradata patterns. Varying levels of interviewing quality also translate into interviewer effects. Interviewing quality is thus conceived to be behind both paradata patterns and interviewer effects. Chapters 2, 3, and 4 devote themselves to exploring the two-way associations between paradata, interviewing quality, and interviewer effects.

Chapter 2 explores associations between paradata patterns and interviewing quality. We first conduct a Principal Components Analysis (PCA) on 10 interview-level paradata measures and, separately, on 8 item-level paradata measures. We then use these principal components to predict either a) the proportion of evaluated items in an interview for which quality flags were raised (interview-level analysis), or b) a binary variable that indicated if a quality flag was raised or not (item-level analysis). We also fit models that use only respondent, interview, and interviewer characteristics; item characteristics are also used in the case of the item-level models. Comparing models that use only paradata inputs, only non-paradata inputs, and full models that include both paradata and non-paradata inputs, can potentially tell us about the source of variation being captured by paradata in predicting interviewing quality.

Chapter 3 investigates associations between interviewing quality and interviewer effects. As stated earlier, to minimize interviewer effects survey organizations train their interviewers to undertake standardized interviewing (Fowler and Mangione 1990) and then track deviations of interviewing behaviors from protocol. But not much is known as to how these deviations are directly associated with interviewer effects. To test this, we first estimate interviewer effects for each of our analysis items using multilevel models, where a vector of respondent characteristics approximates an interpenetrated design (Mahalanobis 1946). Next, we define an interviewer-level 'flag proportion' variable which is the proportion of interviewing quality evaluations for an item for which QC flags were raised. We also define an overall flag proportion variable that is computed across all

evaluated items. These flag proportion variables are then added as inputs to the initial model. The proportion of between-interviewer variance explained by the flag variables is an indicator of the success of the quality process in detecting interviewers contributing to interviewer effects. The performance of the flag variables is benchmarked against other interviewer-level characteristics such as interviewer education levels. We also conduct an observation-level analysis using the interviewing evaluation data to check for possible associations of the substantive outcomes with indicator QC flag variables. Significant coefficients for the QC indicator variables even after controlling for respondent characteristics would indicate that the QC process is detecting possible inaccuracies in estimates due to interviewing issues. Both the interviewer-level and observation-level analyses are repeated for item non-response as well. Here we take advantage of the fact that the outcome variable has the same structure for all variables, i.e, a binary variable indicating either response or non-response. This allows us to pool observations and fit one cross-classified model by incorporating item and respondent random effects in addition to the interviewer random effects.

Based on our findings in Chapters 2 and 3, we use paradata measures to directly predict interviewer effects in Chapter 4. As in Chapter 3, we first estimate interviewer effects for our analysis items. We then use paradata measures defined at the interviewer-level to explain interviewer effects. We fit separate models using the time and non-time paradata measures to assess their relative performance; the literature has focused on time-based paradata measures such as item time but not much is known about non-time paradata measures such as access to help or making remarks. We calibrate the performance of the paradata measures against interviewer-level characteristics and the flag proportions that were defined in Chapter 3. Variable selection for all models is conducted using the ALASSO (adaptive least absolute shrinkage and selection operator). To evaluate our predictions, we undertake a bootstrap-based approach where models are fit to the original data based on variables selected on the resample data.

Each chapter contains practical implications of our findings. Chapter 5 summarizes these implications and also recommends steps for future research.

References

Biemer, P., Herget, D., Morton, J., and Willis, G. (2000). The Feasibility of Monitoring Field Interview Performance Using Computer Audio-recorded Interviewing (CARI). In *Proc. Jt. Stat. Meet. Surv. Res. Methods Sect.*, pages 1068–1073. American Statistical Association.

Biemer, P. and Lyberg, L. (2003). *Introduction to Survey Quality*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Cheung, G., Piskorowski, A., Wood, L., and Peng, H. (2016). Using Survey Paradata. In *IBUC 15th Int. Blaise Users Conf.*, Washington DC, USA.

Couper, M. (1998). Measuring survey quality in a CASIC environment. In *Proc. Jt. Stat. Meet. Am. Stat. Assoc. Surv. Res. Methods Sect.*, pages 41–49, Alexandria, VA. American Statistical Association.

Couper, M. and Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *J. R. Stat. Soc. Ser. A (Statistics Soc.*, 176(1):271–286.

Fowler, F. and Mangione, T. (1990). *Standardized Survey Interviewing: Minimizing Interviewer Related Error*. SAGE Publications, Inc., Thousand Oaks, California.

Groves, R. (1983). Implications of CATI. *Sociol. Methods Res.*, 12(2):199–215.

Groves, R. (1989). *Survey Errors and Survey Costs*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Hox, J. (1998). Multilevel Modeling: When and Why. In Balderjahn, I., Mathar, R., and Schader, M., editors, *Classif. Data Anal. Data Highw. Stud. Classif. Data Anal. Knowl. Organ.*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 147–154. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kish, L. (1962). Studies of Interviewer Variance for Attitudinal Variables. *J. Am. Stat. Assoc.*, 57(297):92.

Maas, C. and Hox, J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology*, 1(3):86–92.

Mahalanobis, P. (1946). Recent Experiments in Statistical Sampling in the Indian Statistical Institute. *J. R. Stat. Soc.*, 109:325–370.

Mitchell, S., Strobl, M., Fahrney, M., Nguyen, T., Bibb, S., Thissen, R., and Stephenson, W. (2008). Using computer audio-recorded interviewing to assess interviewer coding er-

ror. In *Proc. Proc. Jt. Stat. Meet. Surv. Res. Methods Sect.*, Alexandria, VA. American Statistical Association.

Raudenbush, S. (2008). Many Small Groups. In *Handb. Multilevel Anal.*, pages 207–236. Springer, New York, NY.

Schaeffer, N., Dykema, J., and Maynard, D. (2010). Interviewers and Interviewing. In Wright, J. and Marsden, P., editors, *Handb. Surv. Res.*, pages 437–470. Emerald Group Publishing Limited, Bingley, UK, second edi edition.

Steinkamp, S. (1964). The Identification of Effective Interviewers. *J. Am. Stat. Assoc.*, 59(308):1165–1174.

Thissen, R. (2014). Computer Audio-Recorded Interviewing as a Tool for Survey Research. *Soc. Sci. Comput. Rev.*, 32(1):90–104.

Thissen, R., Fisher, C., Barber, L., and Sattalur, S. (2008). Computer Audio-Recorded Interviewing (CARI): A Tool for Monitoring Field Interviewers and Improving Field Data Collection. In *Stat. Canada's Int. Symp. Data Collect. Challenges, Achiev. New Dir.*

Tourangeau, R., Rips, L., and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press, Cambridge.

Tucker, C. (1983). Interviewer Effects in Telephone Surveys. *Public Opin. Q.*, 47(1):84.

West, B. and Blom, A. (2016). Explaining Interviewer Effects: A Research Synthesis. *J. Surv. Stat. Methodol.*

Chapter 2

Can paradata tell us about interviewing quality?

2.1 Introduction

The term paradata was defined by Couper (1998) to refer to keystrokes and time stamps generated during the Computer-Assisted Interviewing (CAI) process. While the term is generally used to refer to any additional data generated in the process of conducting a survey (Kreuter 2013), in this chapter we refer to 'paradata' in the specific sense it was first defined. Paradata have been used in quality control since they are detailed, generated 'free', largely not afflicted by missingness, and relatively error-free (West and Sinibaldi 2013). However, assumptions behind their usage are largely untested. Specifically, little is known about patterns of paradata and their associations with interviewing quality. Figure 2.1 plots two paradata variables - average item time and proportion of interviews in which help was accessed - for each interviewer in the 2015 wave of the Panel Study of Income Dynamics (PSID) for a question on the number of rooms. The figure shows visible between-interviewer differences for item time as well as help access. Moreover, since accessing help takes time, we see a correlation between these two variables. But Figure 2.1 provokes the following questions: Do higher item times indicate better interviewing? Does a higher frequency of help access signify a careful interviewer or does it signify a confused interviewer? Is there a difference in interviewing quality between cases that have the same item time but differ in their access to help?

The literature on this topic is sparse (Couper and Kreuter 2013) and has concentrated on data on timings and their associations with item, respondent, and interviewer characteristics (Couper and Kreuter 2013; Loosveldt and Beullens 2013a,b; Olson and Smyth 2015), instrument design (Couper et al. 1997; Edwards et al. 2008), and 'internal' measures of quality e.g., using rounded responses as a proxy for satisficing (Nix 2014; Turner et al.

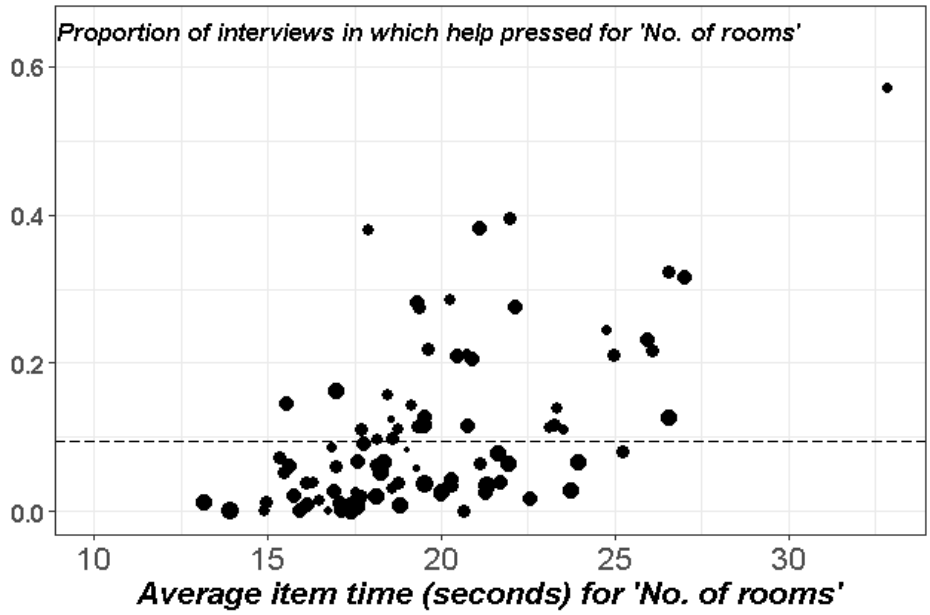

Figure 2.1: Average item time and access to help for item A8 (no. of rooms). Help access is seen to be correlated with item time. Each dot represents one of the 96 PSID 2015 interviewers, with the dot size proportional to interviewer workload; the plot suggests that help access is not correlated with interviewer workload. The horizontal dashed line in the plot is the average proportion of cases for which help was pressed (9.4%).

2015; Vandenplas et al. 2017). The few papers that have looked at multiple paradata variables have focused on their use in survey instrument design and usability training (Couper et al. 1997; Couper 1998; Lepkowski et al. 1998; Hansen et al. 1998; Couper 2000; Bumpstead 2001; Mockovak and Powers 2008). The little research concerning the use of multiple paradata variables for quality control in CAI surveys (Couper 1998; Gu et al. 2013; Joyal 2016) do not seem to have used external (non-paradata based) quality benchmarks.

In the absence of direction from the literature, paradata usage in quality control has been guided by intuitions (Johnson et al. 2001; Penne and Snodgrass 2003; Wang et al. 2013) and rules of thumb (Moshinsky and Carter 2013; Devonshire 2013; Cheung et al. 2016; Hunt 2016) which are largely untested. Finally, since the level of detail in paradata can be overwhelming (Couper et al. 1997; Nicolaas 2011), they are generally aggregated and used in univariate fashion (e.g., Hansen and Marvin 2001).

In summary, studies investigating associations of multivariate paradata patterns with interviewing quality are lacking, a gap that we address in this chapter. If such associations exist, practitioners can utilize them to harness the full range of paradata in an objective way for survey quality control.

We consider both interview-level and item-level inferences. The following are our research questions at the interview-level:

1. What paradata patterns, if any, are associated with interviewing quality?

2. What respondent, interview, and interviewer characteristics - henceforth collectively called non-paradata characteristics - if any, are associated with interviewing quality?

3. How do the paradata and non-paradata variables compare when predicting interviewing quality?

We replicate these questions at the item-level but also additionally include item characteristics as predictors. This chapter is organized as follows: Section 2 gives an overview of the study survey, Section 3 describes the data used for analysis, Section 4 details the analysis methods, Section 5 presents the results obtained, and Section 6 discusses these results and considers how they can be used in practice.

2.2 Study survey

We used data from the 2015 wave of the PSID for our research. The PSID is a nationally representative survey of families and individuals in the U.S., conducted via Computer Assisted Telephone Interviewing (CATI); a small proportion of interviews (approximately 2.5%) are conducted face-to-face but we focus only on data from the CATI interviews. The survey consists of biennial waves where one respondent per family is administered a 'main interview'; supplemental studies are added to this main interview, e.g., the 'transition into adulthood supplement' is asked to individuals when they become 18 years of age. The PSID main interview begins by taking consent from the respondent followed by questions about the family composition and member details. These 'coverscreen' questions are followed by substantive questions. On average, a respondent answers approximately 360 substantive questions from 11 sections as shown in Table 2.1; sections concerning employment (sections BC/DE), expenditures (section F), and health (section H) account for close to 60% of interview duration.

Between March-December 2015, 9048 respondents were interviewed by 96 interviewers with a response rate of 89% (calculated with respect to the previous wave). An interview lasted 80 minutes on average. The detailed questionnaire [1] and codebook [2] are available on the PSID website.

[1] `ftp://ftp.isr.umich.edu/pub/src/psid/questionnaires/q2015.pdf`
[2] `ftp://ftp.isr.umich.edu/pub/src/psid/codebook/fam2015er_codebook.pdf`

No.	Section	Substantive area	Average # items administered in an IW	Average IW duration (mins)
1	A	Housing, utilities, and computer use	36	7
2	BC, DE	Employment	46	22
3	F	Expenditures	52	11
4	G	Current income and other family unit member education	48	9
5	R	Off-year income and public assistance	12	2
6	W	Wealth and active savings	22	5
7	P	Pensions	13	3
8	H	Health	96	14
9	J	Marriages and children	11	1
10	KL	New head and spouse/partner background	14	3
11	M	Philanthropy	9	2
		Average interiew	**358**	**80**

Section BC,DE includes the Event History Calendar (EHC)

Table 2.1: PSID substantive section descriptions.

2.3 Data

To address our research questions we considered five forms of predictor data: paradata - Interview-level; paradata - Item-level; Item characteristics; Respondent and Interview characteristics; and Interviewer characteristics. Data from interviewing quality evaluations form our outcome variable. These data are described below.

2.3.1 Outcome: Interviewing quality evaluations

In 2015, PSID recorded two of the first four interviews in every interviewer's workload followed by a further 10% random sample, resulting in 1120 recorded interviews. A 'capture list' dictated which item in the interview was to be recorded, based on the item's substantive importance. For the first 3 weeks of fieldwork, the capture list inadvertently contained 1157 items belonging to a pretest version. This was corrected and the list pared down to 382 items. We only consider data from the 382 items for our analyses.

Of the 1120 interviews, 594 CATI interviews (53% of all the recorded interviews) were listened to by nine quality control (QC) evaluators. Some of these interviews were randomly chosen while others were subjectively selected; we do not have information on the selection mechanism associated with each interview. Owing to issues such as bad recordings or missing interviewer characteristics, only 555 interviews were available for

analysis. These interviews were conducted by 92 interviewers (96% of the 96 interviewers), with a median of 6 evaluated interviews per interviewer (first quartile: 4 interviews, third quartile: 8 interviews). The recorded items accounted for a median 35% of the total number of administered substantive items (IQR: 31% - 40%) and a median 45% of the substantive interview duration (range: 40% - 50%) within the 555 evaluated interviews.

Apart from their training and extensive experience in behavior coding, many of the QC evaluators have been interviewers themselves which especially equips them to understand interviewer behavior. An evaluator raised a QC flag for an item if she encountered an issue in any of the five interviewing dimensions in Table 2.2.

Table 2.2: The five interviewing evaluation dimensions with sixteen categories.

No.	Interviewing dimension	Categories
1	Question asking	Altered wording; Skipped question; Question delivery; Not verbatim; Other reading error
2	Probing and clarifying	Failure to probe or clarify; Inappropriate, evaluative, or directive probe; Other probing error
3	Data entry	Wrong category; Wrong entry
4	Feedback	Emotive feedback; Other feedback error
5	Other reasons	Unprofessional conduct; Consent error; Household composition; Other error

Coders also explicitly noted the cause of a major flag; four of the sixteen categories accounted for 70% of all major flags: 'failure to probe or clarify' (44%), 'altered wording' (11%), 'inappropriate, evaluative, or directive probe' (9%), and 'other entry error' (6%). The large proportion of flags due to improper probing is expected since interviewers find it the hardest skill to learn (Fowler and Mangione 1990, p.44); Hicks et al. (2010) find that interviewers probed only in 57 percent of the instances when a probe was needed. In line with recommendations in Couper et al. (1992) and Steve et al. (2007, p.404), coders also look for good interviewer behaviors that can be reinforced. Such behaviors are coded as 'positive'. Some coders left the evaluations blank when they encountered satisfactory interviewing, i.e., neither praise-worthy nor flag-worthy, while other coders coded such behavior as 'positive'. For our analyses, we combined the blank and positive evaluations into one 'no flag' category and the major and minor flags into a single 'some flag' category. We had 2329 flags (spread across interviews and interviewers) from 56471 item-by-interview cases. The overall flag rate is therefore 4.1% which varies by interview (IQR: 1.6% - 5.7%) and interviewer (IQR: 2.1% - 5.3%).

2.3.2 Paradata - Interview-level

We defined 10 interview-level measures and computed these from the raw paradata:

1. Count of interview sessions; multiple interview sessions might indicate a time-strapped or difficult respondent. Only sessions greater than 15 minutes were counted

for this measure so as to exclude multiple sessions occurring due to technical issues. Since multiple sessions could take place on the same date, we also included measure 2 below.

2. Range of interview dates (i.e., number of days elapsed between the first and last interview session), top-coded at 30 days.

3. Count of unique items administered; a larger number could impact interviewer and respondent fatigue, potentially associated with more flags.

4. Proportion of items revisited; a low proportion indicates a more straightforward interview. Field visits less than one second were excluded from this measure to filter out transitory item visits.

5. Duration (minutes) spent before the substantive sections of the questionnaire; a large time spent on the coverscreen questions may contribute to compensatory speeding in the substantive part of the questionnaire.

6. Duration (minutes) spent on the substantive part of the interview; this is to distinguish between interviewers who may administer the same number of items (measure 3) but differ in their interviewing pace.

7. Proportion of the interview duration accounted by times up to the first keystroke (computed across items). We use this as a proxy for the amount of time it takes for the interviewer to Ask, Probe and give feedback to the respondent, and Receive a response (abbreviated henceforth as 'APR time'). Behaviors such as quick question delivery and lack of probing ('speeding') would result in a lower magnitude for this measure. A high magnitude for this measure could be a result of longer cognitive processing time by the respondent and/or slower questioning and probing by the interviewer. Only the first item visits were considered to compute this measure since this is when the question would actually have been asked.

8. Proportion of items with remarks; a high proportion could be an indicator of unsure interviewing since interviewers feel the need to justify responses.

9. Proportion of items for which help was accessed; while a high proportion could indicate unsure interviewing, it could also indicate conscientiousness.

10. Count of error messages. During the interviewing process, the CATI software triggers error messages when data which are logically inconsistent or beyond preset numerical ranges are entered. A high count indicates many inconsistent responses which could be due to inadequate interviewing and/or difficult respondents.

Descriptive statistics for these measures are given in Table 2.3. All measures except 'du-

ration before substantive sections' were computed on the substantive sections. Measures 8-10 have small average values but exhibit variation at the upper end of the distributions.

Table 2.3: Descriptive statistics for interview-level paradata measures. These are based on 9048 interviews. Some measures are top-coded as described in Section 2.4.1.

No.	Interview-level measures	Min.	Q1	Median	Mean	Q3	Max.
1	No. of interview sessions	0	1	1	1.2	1	5
2	Range of interview dates	0	0	0	1.5	0	30
3	Number of unique items administered	159	268	346	349	412	709
4	Proportion of items revisited	0	0.04	0.06	0.07	0.09	0.29
5	Interview duration (mins) spent before the substantive sections	0.8	2	4	5	6	33
6	Interview duration (mins) spent on the substantive sections	16	57	71	75	88	200
7	Proportion of interview duration accounted by APR times	0.04	0.08	0.09	0.09	0.12	0.28
8	Proportion of items with remarks	0	0.004	0.009	0.014	0.018	0.14
9	Proportion of items for which help was invoked	0	0	0.003	0.006	0.008	0.08
10	Proportion of error messages	0	0	0	0.0017	0.0025	0.086

2.3.3 Paradata - Item-level

For inferential and computational reasons, we limited ourselves to 171 items that had at least 100 QC evaluations; with small group sizes, the likelihood for higher-level variances can be highly skewed and numeric integration required for our models becomes difficult (Raudenbush 2008, Rabe-Hesketh and Skrondal 2008). We also excluded 'Event History Calendar' which appears in the data as a single 'item' but is actually a complex set of questions. We now had 44927 cases from 170 items for our analysis. These 170 items were distributed across the questionnaire and accounted for a median 35% of the substantive items and substantive interview duration among the evaluated interviews (ranges: 16%-52% and 14%-59% respectively).

The item-level dataset had 1.03 million item-interview rows each associated with 8 measures that we computed from the raw paradata. These item-level measures are largely analogous to the interview-level measures and are as follows:

1. Count of multiple item visits. Item visits less than 1 second were excluded from the computing since these could just be transitory visits.
2. Item time on the first visit.
3. APR time on the first visit.
4. Keycounts.
5. Count of mouse clicks.
6. Count of the number of times the remark option was invoked.

7. Count of the number of times help was accessed.

8. Count of error messages.

Descriptive statistics for these measures are given in Table 2.4. While many measures are sparse, there is substantial item-level variation.

Table 2.4: Descriptive statistics for the item-level paradata measures. These are based on 1.03 million item-within-interview observations. Some measures are top-coded as described in Section 2.4.1.

No.	Item-level measures	Min.	Q1	Median	Mean	Q3	Max.
1	Count of multiple item visits	0	0	0	0.1	0	3
2	Item time (seconds)	0	5	9	12	14	822
3	APR time (seconds)	0	4	7.5	9.7	12	281
4	Keycounts	0	2	2	5	3	184
5	Count of mouseclicks	0	0	0	0.14	0	20
6	Count of times remark invoked	0	0	0	0.02	0	3
7	Count of times help accessed	0	0	0	0.01	0	2
8	Count of error messages	0	0	0	0	0	2

2.3.4 Item characteristics

For each of the 170 items, we coded five variables as follows:

1. Whether an item may be considered sensitive (30 items) or not (140 items); sensitive items may be more subject to quality issues. To classify an item as sensitive, we followed the guidance in Tourangeau et al. (2000) and Tourangeau and Yan (2007) and gauged whether it was one or more of the following: could be perceived by the respondent as intrusive, could raise concerns of disclosure risk, or could evoke socially desirable responses, e.g., item A27G asks how likely is it that the respondent will continue to be behind on their mortgage/loan payments in the next 12 months.

2. A variable called 'RecallHeavy' that checks if an item relied heavily on the respondent's memory and/or would be likely to require reference or very specific knowledge (51 items) or not (119 items), e.g., item A21 asks for 'yearly property taxes, including city, county, and school taxes' or item A24 asks for the remaining principal on the housing loan). The interviewer's task is more difficult for such items since they might require more probing.

3. Whether the item had a specific probing instruction to the interviewer (22 items) or not (148 items), e.g., item A31 is a question on monthly rent. It has a specific

instruction to the interviewer to probe if the given response is only that family's share, in case the dwelling is shared by more than one family.

4. Whether the question had any special instruction (apart from any probing instruction) to the interviewer (36 items) or not (134 items), e.g., in the case of item A31 mentioned above, if the dwelling is a mobile home, then the instruction is: "If family unit owns the lot, do not include the value of the lot".

5. Response type was coded into 7 categories as follows: binary (74 items), multinomial with an 'other-specify' option (27 items), multinomial but no 'other-specify' option (10 items), numeric monetary values (35 items, e.g., income), numeric non-monetary values (14 items, e.g., number of rooms), open-ended (4 items), and others (6 items, e.g., item F49b on vehicle type where the instrument has a drop-down list to choose from). The modal binary response category was chosen as the reference category. The 'numeric monetary' category includes not only $ value responses but also responses on interest rates.

A list of the 170 items and the above associated variables is in Appendix 2.A.

2.3.5 Respondent and interview characteristics

We included eight respondent and interview characteristics in our analyses as follows:

1. Respondent sex (female respondents: 60% among the evaluated interviews); past research suggests that respondent sex may be associated with response effects (Skowronski and Thompson 1990; Auriat 1993; Lee and Lee 2012).

2. Respondent education in years (mean: 13.5 years, standard deviation: 2.3 years); a measure of cognitive sophistication (Krosnick and Alwin 1987; Groves 1989; Krosnick 1991; Knauper 1999).

3. Respondent age in years (mean: 50 years, standard deviation: 17 years); this variable is known to be correlated with response errors even after accounting for education (Fowler and Mangione 1990; Knauper 1999).

4. Number of waves as a respondent in the last five waves; more frequent respondents may pose fewer issues to the interviewer. Seventy-eight percent of respondents in the evaluated interviews were present for at least 4 waves.

5. Number of adults in the family unit (mean: 1.7, standard deviation: 0.8); higher values are potentially more burdensome to the respondent and interviewer since some items must be repeated for every adult.

6. Number of calls exchanged between the respondent and interviewer prior to the interview (mean: 14.4, standard deviation: 22.7); a difficult-to-reach respondent may also be pressed for time, leading to speeding by the interviewer.

7. A binary variable based on interviewer-provided information that indicated if the respondent had records, statements, or other documents readily available for reference during the interview; respondents with documents available may be less likely to need aiding by interviewers, reducing the probability of a quality flag occurring. Fourteen percent of respondents were reported to have these available.

8. A binary variable that indicated whether the interview was the first two in the interviewer's sequence (10% of the evaluated interviews were in this category) or not. The initial interview pair may be associated with a higher flag proportion as interviewers start to get accustomed to the wave.

2.3.6 Interviewer characteristics

We used 3 demographic variables and 3 variables derived from interviewers' work characteristics.

1. Interviewer sex (88% female).

2. Interviewer age (mean: 53.6 years, standard deviation: 12.1 years).

3. Interviewer education, which was categorized as: less than High school (12% of interviewers), high school/GED (35% of interviewers), some college (28% of interviewers), and college graduate and above (25% of interviewers). The high school/GED category was used as the reference category for our analyses.

4. Interviewer workload, i.e., number of conducted interviews (mean: 114.5 interviews, standard deviation: 41.6 interviews)

5. Mean interviews per day (mean: 1.2 interviews, standard deviation: 0.12 interviews); even a moderate workload may lead to interviewer fatigue if completed in a short time period.

6. The coefficient of variation of the number of daily interviews conducted (mean: 0.53, standard deviation: 0.11 interviews); interviewers with more consistent daily workloads may be more organized interviewers associated with better interviewing quality.

2.4 Methods

2.4.1 Principal components analysis (PCA)

Despite efforts to create independent paradata measures, some of them were correlated, e.g., item counts and interview duration, leading to multicollinearity issues in our models. Also, we were more interested in *patterns* of paradata. Therefore, we turned to Principal Components Analysis (PCA, Jolliffe 1986). We conducted two PCAs - one each at the interview-level and item-level - using the correlation matrices of the respective centered and scaled paradata measures. PCA performs an orthogonal transformation of the data such that the same number of dimensions (Principal Components, PCs) are returned as the number of paradata measures, but these new dimensions are now linearly uncorrelated. The analysis also returns a rotation matrix which is a matrix of correlations ('loadings') of each paradata measure with the PCs, thus allowing us to interpret the PCs. Each interview (in the interview-level analysis) or observation (in the item-level analysis) now has a value for each PC, called a 'score'. We will use these scores as inputs in our models.

Even though our models can ultimately use data from only the 555 evaluated interviews, we did not want to limit our understanding of paradata patterns to these interviews. Therefore, we ran the PCA on all interviews and used scores from the subset of the 555 evaluated interviews for modeling. Generally, in PCA, interest is in the first few large components. However, in a regression context, PCA does not involve the outcome variable; substantively interesting small components may be associated with the outcome variable but the big components may not have any associations (Hadi and Ling 1998; Faraway 2005, p.144). We therefore extracted PCs that cumulatively account for 90% variation in the paradata measures, resulting in the extraction of 8 interview-level PCs and 6 item-level PCs.

Some paradata were extreme and clearly the result of technical issues, e.g., an interview duration of more than 37 hours. Moreover, PCA is sensitive to extreme observations and we did not want a few cases, even if genuine, to reduce the generalizability of our analysis. On the other hand, we wanted to be conservative since observations towards the tails can convey valuable information. After some sensitivity analyses, we top-coded the raw time-based paradata variables at the 99.95^{th} percentile and used these to compute our measures. Other paradata measures were visually inspected and extreme data points top-coded. The PCA analyses were conducted in R (Team 2013) using the *prcomp* function.

2.4.2 Modeling

Interview-level model

Our outcome variable is $Y_{ij} \sim BIN(N_{ij}, p_{ij})$,where Y_{ij} and N_{ij} are the number of flags and the number of items respectively, assessed in interview $j = 1, 2, \ldots, n_i$ conducted by interviewer $i = 1, 2, \ldots, 92$. We used interview-level PC scores as inputs in the following varying interviewer-intercept logistic model, where $x_{c_I ij}$ is the ${c_I}^{th}$ PC score (the 'I' in the subscript indicates that this is an interview-level score). Initial models that we fit showed signs of overdispersion with the dispersion parameter being approximately 1.31, computed using the *dispersion_glmer* function in the *blmeco* package (Korner-Nievergelt et al. 2015) in R (Team 2013). We therefore introduce an Observation Level Random Effect (OLRE), denoted by δ_{ij} in the model below, to correct for this (Browne et al. 2005; Skrondal and Rabe-Hesketh 2007).

$$\begin{aligned}\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_o + \delta_{ij} + u_i \; + \sum_{c_I=1}^{8} \beta_{c_I} x_{c_I ij} + \\ \beta_9 RespMale_j + \beta_{10} RespAge_j + \beta_{11} RespEducation_j + \\ \beta_{12} TimesRespLast5Waves_j + \beta_{13} AdultsFamilyUnit_j + \\ \beta_{14} Calls_j + \beta_{15} IWseqFirstTwo_j + \beta_{16} ReferenceDocs_j + \\ \beta_{17} IwerMale_i + \beta_{18} IwerAge_i + \beta_{19} IwerEduc\_LessThanHS_i + \\ \beta_{20} IwerEduc\_SomeCollege_i \; + \beta_{21} IwerEduc\_Graduate_i + \\ \beta_{22} IwerWorkload_i + \beta_{23} IwerMeanDailyIW_i \; + \beta_{24} IwerCVDailyIW_i \end{aligned} \tag{2.1}$$

$$u_i \sim N(0, \sigma^2_{iwer}) \tag{2.2}$$

$$\delta_{ij} \sim N(0, \sigma^2_{olre}) \tag{2.3}$$

We tried introducing interactions, especially cross-level interactions between respondent and interviewer sex, age, and education, and between the variable indicating the first 2 interviews and the interviewer workload variables. However, none of these were found to be significant and were dropped from the model. Before fitting the full model above, we fit the following subset models: a model with only the PC terms ('Paradata model') and a model with only the non-paradata variables ('Non-paradata model'). Both these

models also accounted for overdispersion, the dispersion parameters being 1.34 and 1.36 respectively. The full model was constructed by adding terms from these subset models.

Item-level analysis

Our outcome here is a Bernoulli variable, $Y_{ijk} \sim BER(p_{ijk})$, set equal to 1 if a QC flag occurs for item $k = 1, 2, \ldots, 170$ within interview $j = 1, 2, \ldots, n_i$ administered by interviewer $i = 1, 2, \ldots, 92$. While an interview is nested within an interviewer, an item occurs across multiple interviews resulting in a cross-classified data structure allowing us to specify random item intercepts u_k in addition to random interview intercepts u_j and random interviewer intercepts u'_i. The full model, which includes interactions between the item characteristics and the paradata and non-paradata terms, is as follows, where where x_{cijk} is the c^{th} PC score:

$$\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \beta_o + u'_i + u_j + u_k \; + \sum_{c=1}^{6} \beta_c x_{cijk} +$$

$$\beta_7 Sensitive_k + \beta_8 RecallHeavy_k + \beta_9 ProbingInstruc_k \; + \beta_{10} SpecialInstruc_k +$$

$$\beta_{11} ResponseType\_NumericMonetary_k +$$

$$\beta_{12} ResponseType\_NumericNonMonetary_k +$$

$$\beta_{13} ResponseType\_MultinomialOther_k +$$

$$\beta_{14} ResponseType\_MultinomialNoOther_k +$$

$$\beta_{15} ResponseType\_OpenEnded_k \; + \beta_{16} ResponseType\_Others_k +$$

$$\beta_{17} RespMale_j + \beta_{18} RespAge_j + \beta_{19} RespEducation_j +$$

$$\beta_{20} TimesRespLast5Waves_j + \beta_{21} AdultsFamilyUnit_j +$$

$$\beta_{22} Calls_j + \beta_{23} IWseqFirstTwo_j + \beta_{24} ReferenceDocs_j +$$

$$\beta_{25} IwerMale_i + \beta_{26} IwerAge_i + \beta_{27} IwerEduc\_LessThanHS_i +$$

$$\beta_{28} IwerEduc\_SomeCollege_i \; + \beta_{29} IwerEduc\_Graduate_i +$$

$$\beta_{30} IwerWorkload_i + \beta_{31} IwerMeanDailyIW_i \; + \beta_{32} IwerCVDailyIW_i +$$

$$\beta_{33}(x_{1ijk} * Sensitive_k) \; + \beta_{34}(x_{1ijk} * ProbingInstruc_k) \; + \beta_{35}(x_{1ijk} * SpecialInstruc_k) +$$

$$\beta_{36}(x_{2ijk} * ResponseType\_NumericMonetary_k) +$$

$$\beta_{37}(x_{2ijk} * ResponseType\_NumericNonMonetary_k) +$$

$$\beta_{38}(x_{2ijk} * ResponseType\_MultinomialOther_k) +$$

$$\beta_{39}(x_{2ijk} * ResponseType\_MultinomialNoOther_k) +$$

$$\beta_{40}(x_{2ijk} * ResponseType\_OpenEnded_k) +$$

$$\beta_{41}(x_{2ijk} * ResponseType\_Others_k) +$$

$$\beta_{42}(x_{3ijk} * RecallHeavy_k) \; + \beta_{43}(x_{5ijk} * RecallHeavy_k) \; + \beta_{44}(x_{6ijk} * SplInstruc_k) \; +$$

$$\beta_{45}(RecallHeavy_k * IwerEduc\_LessThanHS_i \;) +$$

$$\beta_{46}(RecallHeavy_k * IwerEduc\_SomeCollege_i \;) +$$

$$\beta_{47}(RecallHeavy_k * IwerEduc\_Graduate_i \;) \tag{2.4}$$

$$(u_i', u_j, u_k) \overset{iid}{\sim} N(\mathbf{0}, \mathbf{D}) \tag{2.5}$$

$$\mathbf{D} = \begin{bmatrix} \sigma_{iwer}'^2 & 0 & 0 \\ 0 & \sigma_{iw}^2 & 0 \\ 0 & 0 & \sigma_{item}^2 \end{bmatrix} \tag{2.6}$$

Similar to the interview-level analysis, we preceded the fitting of the full model by fitting subset models. These are as follows:

1. A model with only the item characteristic terms ('Item model').
2. A model with only the paradata terms ('Paradata model).
3. A model that additively includes terms from the above 2 models ('Item + Paradata' model). Comparing coefficients from this model to the above 2 models can tell us about common information between the paradata and non-paradata variables in predicting occurrence of a flag.
4. A model that also allows for interactions between the item characteristic and PC terms ('Item x Paradata' model).

Fitting of subset models was also conducted for the non-paradata variables as follows:

5. A model with only the non-paradata terms ('Non-paradata' model).
6. A model that additively includes the item characteristic and paradata terms ('Item + Non-paradata' model).
7. A model that also allows for interactions between the item characteristic and non-paradata terms ('Item x Non-paradata' model).

The full model includes terms from the 'Item x Paradata' model and 'Item x Non-paradata' model.

2.4.3 Model fitting and inference

Before model fitting, we looked at the pairwise correlations among PC scores (for the subset of interviews that we used for modeling) and visually inspected the distribution of scores; we did not find any problem. After this, we centered and scaled numeric model inputs. All multilevel models were fit with the *lme4* package (Bates et al. 2015) using the Laplace approximation in the R (Team 2013) software. Logistic models in *lme4* use

ML estimation. Rather than use the default BOBYQA optimizer, we used the 'NLopt' implementation of the BOBQYA optimizer (Johnson) via its R interface 'nloptr' (Ypma et al. 2014); some testing showed that estimates were almost exactly the same as when the default optimizer was used but with more than a 50% reduction in runtime.

In the interview-level case, we computed Cook's distances and DFBETAs and found two interviews with large magnitudes of these statistics which were impacting effect sizes and significance of our estimates. These interviews were dropped and the model re-run with the remaining 553 interviews. We did not compute the leave-one-out diagnostic measures for the item-level model due to the size of the dataset.

We assess the fit of our models using 2 methods.

1. Model predictions are compared to observed data. While this will be anti-conservative since we are using the same data for fitting and prediction, it gives us an approximate measure of the predictive utility of our models. For the item-level model, we also conducted an ROC curve analysis using the *pROC* package (Robin et al. 2011) in R.

2. We undertake simulation-based diagnostics described by Hartig (2018). The key idea is that data simulated from the fitted model should mimic the observed data if the fitted model was correctly specified (Gelman and Hill 2006, p. 158-159). To do this, a thousand datasets are simulated from the model, conditioning on all random effects. Then, for each observation a quantile residual (Dunn and Smyth 1996) - defined as the proportion of simulated values larger than the observed value - is computed and two plots are constructed as described below.

    - If there are no model fit issues, we would expect the quantile residuals across observations to be uniformly distributed. We draw a quantile-quantile plot to evaluate this; more formally, a Kolmogorov-Smirnov test is conducted to detect deviation from uniformity.

    - The quantile residuals are plotted against the mean simulated value for each observation (similar to the diagnostic plot of residuals versus fitted values constructed for a linear model). The mean simulated values are rank transformed and scaled to make it easier to spot issues. To make the analysis more concrete and help protect against missing patterns visually (especially when there are a lot of observations such as in the item-level model), a quantile regression is conducted between the 25th percentile, median, and 75th percentile of the mean simulated values and the quantile residuals; the quantile regression lines should ideally match horizontal lines at these percentiles that would indicate no association between the residuals and the mean simulated responses.

The quantile regression is conducted using quantile regression neural network models via the QRNN package (Cannon 2011) in R, so as to be able to spot potential non-linearities in the patterns.

R code for undertaking the simulation-based analyses was adapted from the source code of the DHARMa package (Hartig 2018) and is included in Appendix 2.B.

Our analysis involves comparing coefficients across models. However, for logistic models the variance at the lowest-level is fixed at $\pi^2/3$ which means that each time covariates are included at this level, the latent variable distribution underlying the logistic distribution is rescaled to be able to hold this variance constant. This scale change tends to inflate the regression coefficients (Snijders and Bosker 1999, p.228-229) and "makes it impossible to compare regression coefficients across models, or to investigate how variance components change" (Hox 2010, p.134) since we cannot separate the impact of scale changes from real substantive changes (Austin and Merlo 2017). To be able to compare results across models, we follow the procedure given in Hox 2010 (p.136) by scaling our variance components by the ratio of the variance under the null model (no covariates used) to variance under the fitted model; regression coefficients were scaled by the square root of this ratio.

We accounted for multiple comparisons in our inferences by adjusting p-values and confidence intervals using the Benjamini-Hochberg (B-H) False Discovery Rate (FDR) method (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2005).

2.5 Results

2.5.1 Interview-level analysis

PCA

Table 2.5 shows the loading structure for the first 8 PCs that were extracted. For clearer interpretation, only loadings that are more than 0.25 are shown in the table. The PCs make for coherent interpretations, e.g., interviews with high scores on PC6 are those which exhibit speeding as well as apparently compensatory item revisits. Short PC labels based on the loadings are included in Table 2.5. We see that PCs with smaller proportions of variance explained in Table 2.5 are also more sharply defined, tending to strongly correlate on fewer measures.

Table 2.5: Interview-level PCA results. Columns are arranged in descending order of the proportion of variation explained. Data (apart from the first two rows) are loadings, i.e., correlations. Unless they aid interpretation, only loadings with an absolute value ≥ 0.25 are shown. Shaded columns are PCs that have significant model coefficients in the paradata model with the '+' indicating an increase in the odds ratio of a flag due to the PC and a '-' indicating a decrease in the odds ratio of a flag.

	+	+		+	+	+		-
	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
Proportion of variance explained	0.26	0.19	0.13	0.10	0.08	0.07	0.06	0.06
Cumulative proportion of variance explained	0.26	0.45	0.58	0.67	0.76	0.83	0.89	0.95
PC label	Unsure long multi-session interviews with multiple item visits	Unsure interviews	Multi-day interviews	Error-message prone interviews	Short-coverscreen interviews	Speeding interviews with multiple item visits	Slow-paced interviewing also involving multiple item visits	Interviews invoking help (but not remarks)
Interview-level paradata measures								
1. No. of interview sessions	0.35		0.47					
2. Range of interview dates	0.26		0.62					
3. Number of unique items administered	0.26	-0.50	-0.46					
4. Proportion of items revisited	0.34					0.78	0.37	
5. Interview duration spent before the substantive sections	0.27				-0.89			
6. Interview duration spent on the substantive sections	0.49							
7. Proportion of interview duration accounted by APR times	0.36	0.37				-0.47	0.56	
8. Proportion of items with remarks	0.29	0.43					-0.50	-0.67
9. Proportion of items for which help invoked	0.29	0.39					-0.43	0.66
10. Proportion of items with error messages				0.93				

Model results: Coefficients

Figure 2.2 plots the odds ratios from the paradata model (dashed lines) and paradata terms from the full model (solid lines). The confidence intervals (CIs) were computed by exponentiating the endpoints of the FDR-adjusted log-odds CIs rather than directly using the standard errors on the odds ratio scale, thus giving us better coverage (Faraway 2006). The scaling factors for the paradata and full models were 0.93 and 0.94 respectively. A detailed table containing the coefficient estimates, standard errors and p-values for the interview-level analyses is in Appendix 2.C.

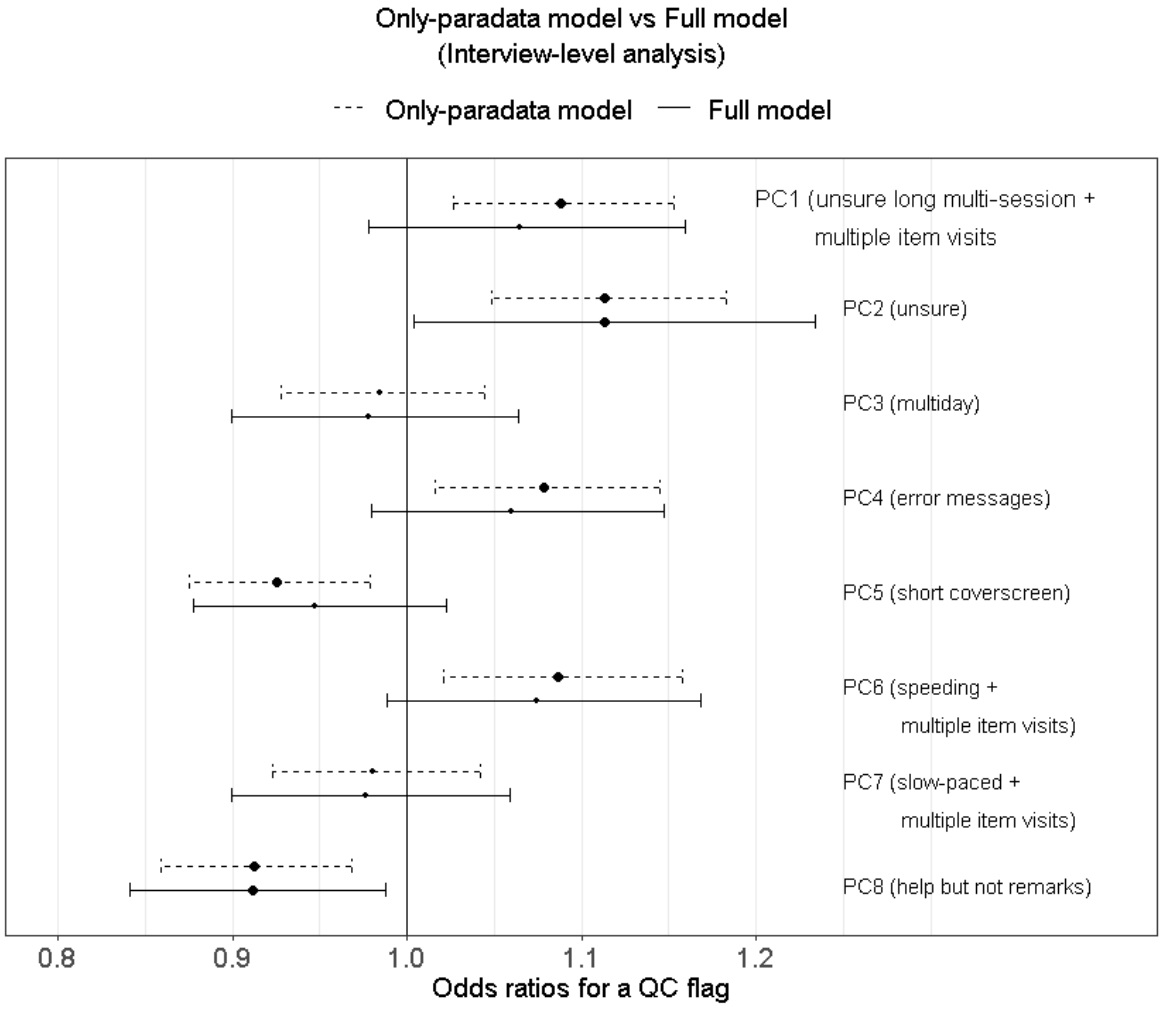

Figure 2.2: Interview-level analysis: Only-paradata model versus Full model. The odds ratios for the only-paradata model terms are shown by dashed lines and those for the fully-adjusted model are shown by solid lines. These effects correspond to scaled PC scores. The horizontal bars are B-H adjusted 90% confidence intervals. The intercept has not been shown in these plots. Estimates whose confidence intervals do not contain 1 are shown with a bigger point size.

We first focus on the paradata model and find 6 of the 8 PCs with significant effects. Interviews with positive scores on 4 PCs - PC1 (unsure long multi-session interviews with multiple item visits), PC2 (unsure interviews), PC4 (error-prone interviews), and

PC6 (speeding interviews with multiple item visits) - are associated with a statistically significant increase in the odds ratio of a flag; for every 1 standard deviation increase in the scores for these PCs, there is an 8% to 11% increase (depending on the PC) in the odds ratio of a flag. Two PCs - PC5 (Short-coverscreen interviews) and PC8 (interviews invoking help but not remarks) are associated with a decrease in odds ratio of a flag occurring; odds ratios for the 2 PCs are 0.93 and 0.91 respectively.

From these results, we see that relying only on univariate paradata measures may be misleading, e.g., interviews that have positive PC2 (unsure interviews) scores are those which seem to spend more time administering items on average and with fewer items to administer. However they are still predicted to have a higher odds ratio of a QC flag. This may be due to their relatively higher help-access and remark-making rates that not only indicate unsure interviewing by themselves, but could also be behaviors that disrupt the flow of an interview leading to more errors. Another example lies in the comparison of PC6 (speeding interviews with multiple item visits) and PC7 (slow paced interviewing also involving multiple item visits). The 'multiple item visit' measure loads on both these PCs with a fairly high magnitude. But in PC7's case, this is also combined with higher APR times and lesser occurrence of help access and remark making. These mixed behaviors likely go in opposite effect directions leading to no significant PC7 effect. However, PC6 is strongly identified with only multiple item visits leading to an increased odds ratio of a flag.

We now examine relationships for the PC terms in the full model, shown by solid lines in Figure 2.2. Of the 6 PCs that were significant in the paradata model, only 2 remain significant in the full model. This means that behaviors that were predictive of interviewing quality and captured by these 4 PCs - PC1 (unsure long multi-session interviews with multiple item visits), PC4 (error-prone interviews), PC5 (short-coverscreen interviews), and PC6 (speeding interviews with multiple item visits) - have their origin in respondent, interview, and interviewer characteristics. The 2 PCs that still remain significant are PC2 (unsure interviews) and PC8 (interviews invoking help but not remarks). Behaviors represented by these PCs possibly originate from interviewing idiosyncrasies that are at least partially independent of respondent, interview, and interviewer characteristics in predicting interviewing quality. The loss of significance of the 4 PCs could simply be because of reduced degrees of freedom since we are adding several more terms in the full model. But we also generally find that the effect sizes of these PCs has changed; in contrast the 2 PCs that still remain significant in the full model have their effect magnitudes more or less unchanged. This lends more credence to our inferences. We follow the above line of analysis for the non-paradata terms as well (scaling factor for the non-paradata model was 0.93). Figure 2.3 shows that of the 8 respondent and interview characteristics, 4 relationships are significant in the non-paradata model. Compared to

female respondents, male respondents on average have a 15% lesser odds ratio of a flag occurring. Every standard deviation increase in education years over the mean reduces the odds ratio of a flag by 8%. On the other hand, every standard deviation increase (17 years) in respondent age over the mean (50 years) increases the odds ratio of a flag by 8%. This means that interviewers, in general, find interviewing more educated respondents less challenging and older respondents more challenging. On average, the first two interviews in an interviewer's workload are associated with a substantially higher odds ratio of a flag (29%) compared to other interviews. This finding suggests that early interviews - not limited to only the 1st two interviews - need to be monitored carefully. In the full model, only this particular term among the respondent and interview variables remains significant. Of the 3 other terms that became insignificant we see a fair reduction in the effect size as well.

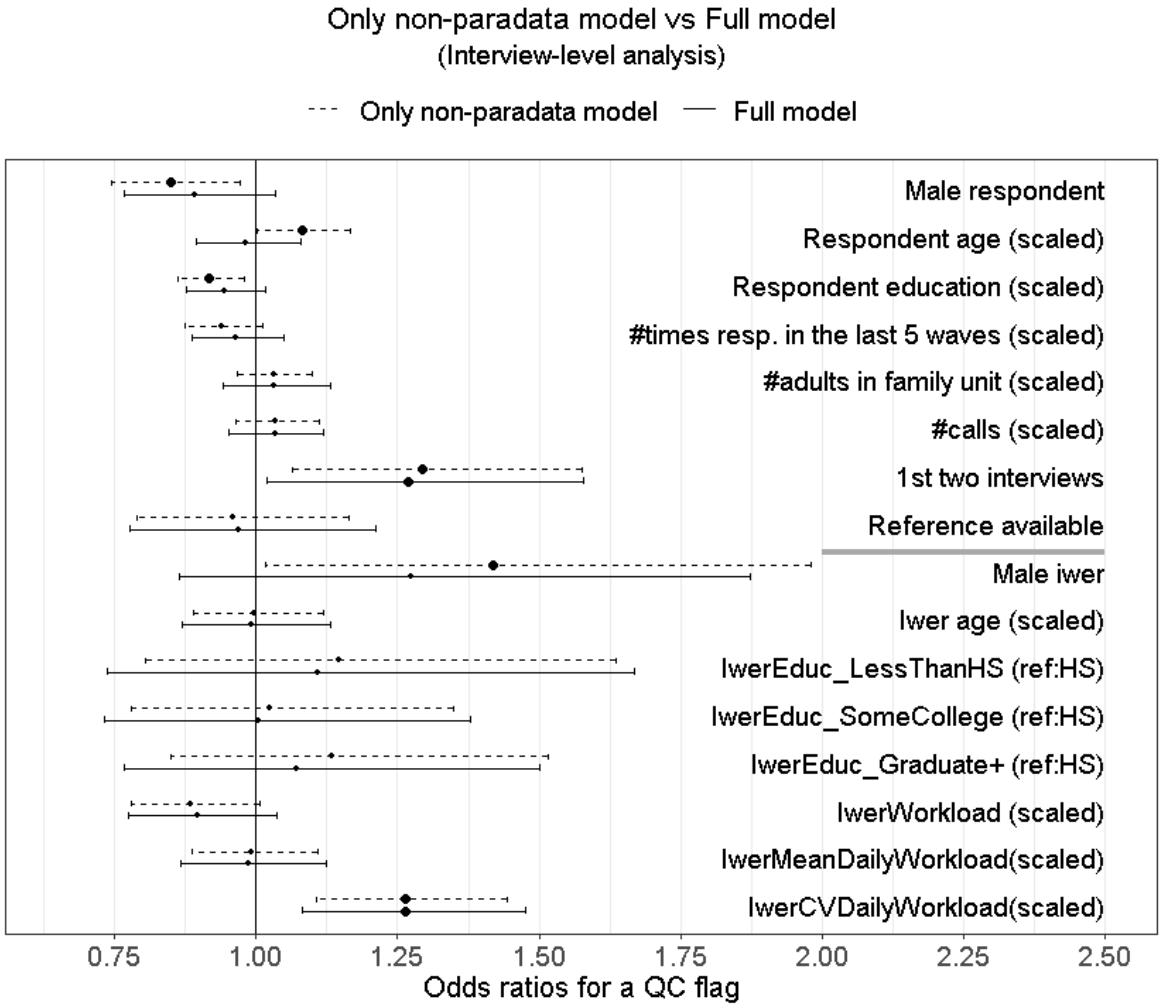

Figure 2.3: Interview-level analysis: Only non-paradata model versus Full model. The odds ratios for the only non-paradata model are shown by dashed lines and those from the fully-adjusted model are shown by solid lines. The horizontal bars are B-H adjusted 90% confidence intervals. The intercept has not been shown in these plots. Estimates whose confidence intervals do not contain 1 are shown with a bigger point size. The thick grey line on the right of the plot separates the respondent and interview terms from the interviewer terms. 'HS' in the interviewer education labels stands for 'high school'.

Of the 8 interviewer characteristics, only 2 have significant relationships in the non-paradata model. Compared to female interviewers, male interviewers on average are associated with a 42% higher odds ratio of a flag occurring. The wide confidence intervals for this effect are due to the small number of male interviewers in the study. After controlling for the PC scores, however, the odds ratio for this term reduces to 1.27 and becomes insignificant. The only interviewer term that is significant in both the non-paradata and full models is the CV of daily workload; the effect sizes for this term are the same in both models. The non-paradata analysis results again demonstrate that there is common information between the PC scores and the respondent, interview, and interviewer characteristics in predicting interviewing quality.

Model Results: Interviewer variance, model fit, and predictive power

Table 2.6 shows the interviewer variance component and model fit indicators for the 3 models. The variances were tested using a 50:50 χ^2 approach (Self and Liang 1987) and were all significant. The scaling factors for the variance terms were 0.87, 0.88, and 0.89 for the paradata, non-paradata, and full models respectively. Even though the paradata model did not have explicit interviewer-level terms, it still manages to account for approximately a third of the interviewer variance. This can happen only when there are between-interviewer differences in mean PC scores, demonstrating that paradata are capturing differences in interviewer behaviors related to interviewing quality. In terms of the AIC and BIC indicators, the paradata model performs the best. All 3 models were able to predict the proportion of flags within ±2% for at least 61% of the interviews.

Table 2.6: Interview-level analysis : Interviewer variance and model fit summary

		Paradata model	Non-paradata model	Full model
(Null model $\hat{\sigma}^2_{iwer} = 0.28$)	$\hat{\sigma}^2_{iwer}$	0.19	0.14	0.15
	AIC	2588	2607	2591
	BIC	2635	2689	2708
Interviews with predicted flag proportion within ±2% of the actual proportion		61%	62%	63%

We assess model predictions in more detail in Figure 2.4. Here, we categorize the actual flag proportions into deciles. Then, within each decile we plot the mean actual proportion of flags and mean predicted proportion of flags for the 3 models. In conducting the predictions, we conditioned on specific interviewers which is reasonable since a large majority of interviewers tend to be used across PSID waves. The predictions are not very good, especially at the tails of the distribution. That considered, the paradata model does better than the non-paradata model at the upper tail.

Figure 2.5 displays the goodness-of-fit plot. The left panel shows that there is a deviation of the quantile residuals from the expected uniform distribution which is also confirmed by the Kolmogorov-Smirnov test. The right panel shows that while the quantile regression line corresponding to the median is close to the expected line, the 25th percentile line shows a quadratic nature indicating that it might be worth considering quadratic transformations of the predictors. But this would also make interpretation a little more difficult.

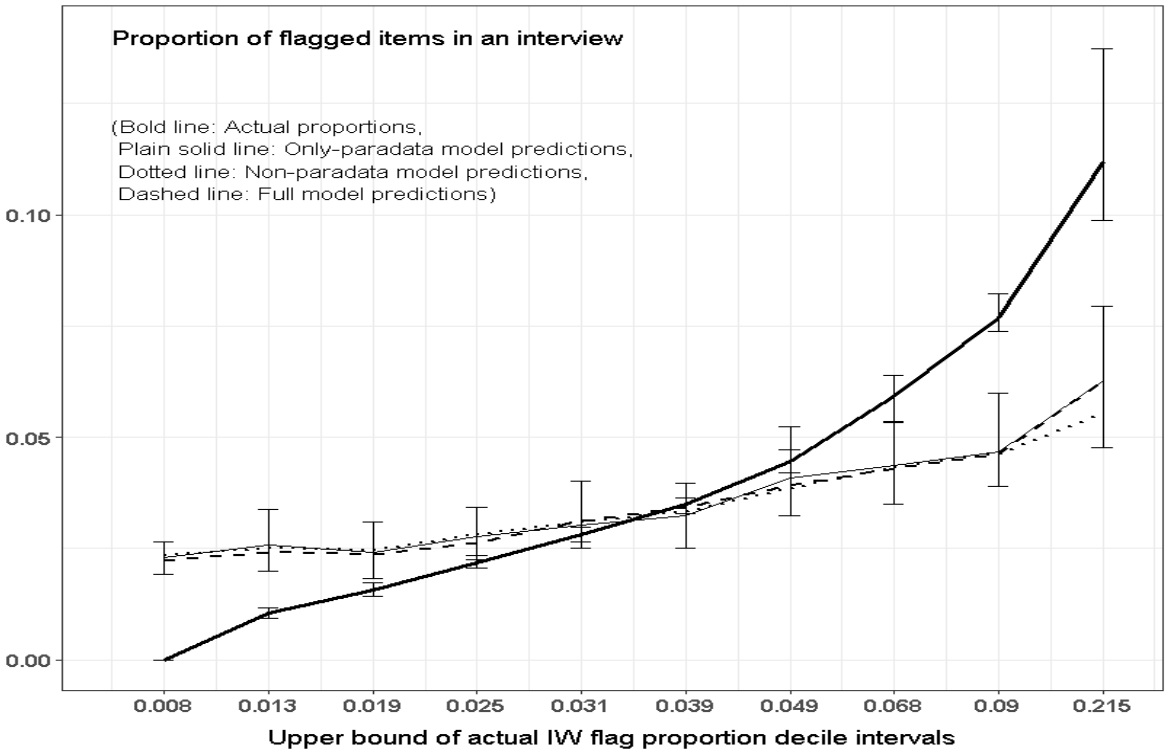

Figure 2.4: Interview-level analysis: Comparison of model predictions. Mean *predicted* flag proportions of the three models are compared to the mean *actual* flag proportions in each of the actual flag proportion deciles. The vertical bars are the IQRs for the actual flag proportions and full model flag proportion predictions.

2.5.2 Item-level analysis

PCA

Table 2.7 displays the loading structure for the 6 item-level paradata PCAs that were extracted. Similar to the interview-level results, all PCs are interpretable and corresponding labels are given in Table 2.7, e.g., we label PC1 as 'unsure' due to the presence of remarks as well as higher mouseclicks without a noticeable loading for multiple visits, indicating that response edits might be taking place on the original item visit. We do not use the term 'speeding' here in describing the PCs since short APR times could simply

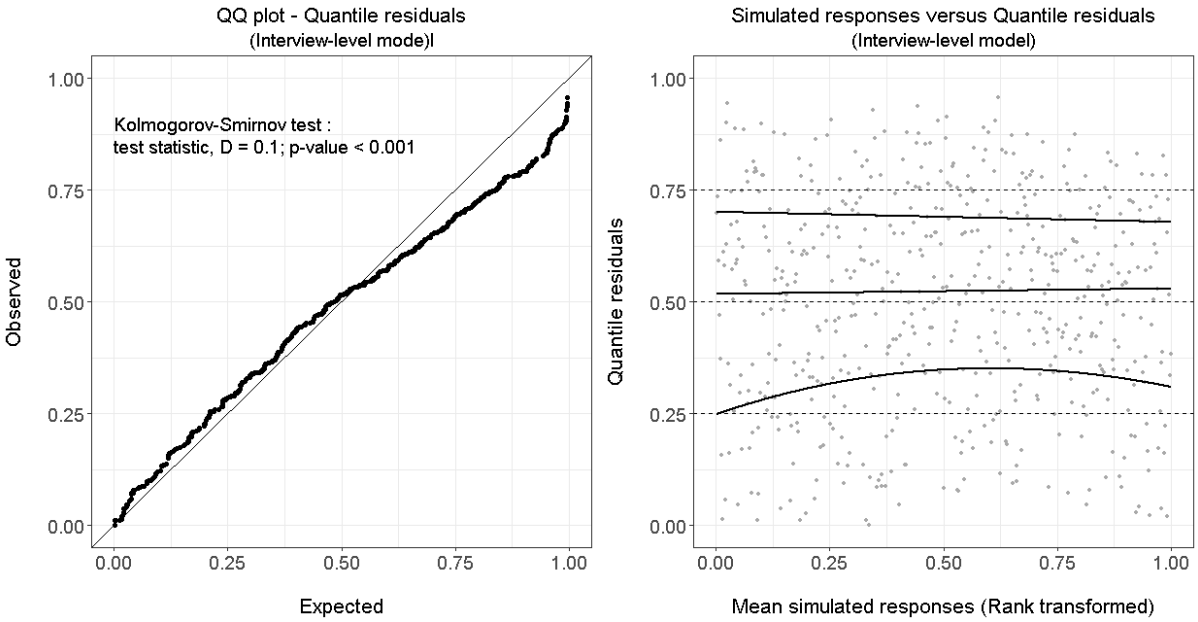

Figure 2.5: Interview-level model diagnostics. The left panel compares the quantile residuals to draws from a uniform distribution. Each point in the right panel is the mean simulated response (across 1000 simulations) for an observation in the data (there are therefore 553 points in this panel). The solid lines in the correspond to the quantile regression lines and the dotted lines are benchmarks for these lines.

be a function of the specific items driving that PC. In contrast to interview-level PCs, we see that even PCs with smaller proportions of variance are correlated with several variables with reasonable magnitude, due to the heterogeneity introduced by the items.

Table 2.7: Item-level PCA results. Columns are arranged in descending order of the proportion of variation explained. Data (apart from the first two rows) are loadings, i.e., correlations. Unless they aid interpretation, only loadings with an absolute value ≥ 0.25 are shown. Key loadings are in bold. Unlike the interview-level PCA results, columns are unshaded since all PC coefficients were significant in the item-level paradata model.

	PC1	PC2	PC3	PC4	PC5	PC6
Proportion of variance explained	0.30	0.16	0.13	0.13	0.10	0.09
Cumulative proportion of variance explained	0.30	0.46	0.59	0.72	0.83	0.92
PC label	Unsure	Multiple visit cases with short remarks and short APR times	Single-visit cases without help and short APR times	Error message-prone cases	Single-visit error-prone cases using help	Remark-prone single-visit cases
Item-level paradata measures						
Multiple visit count		**0.49**	**-0.44**		**-0.64**	**-0.36**
Item time	**0.55**	-0.37				
APR time	0.29	**-0.58**	**-0.37**	0.21	-0.28	0.25
Keycounts	**0.52**		0.29			-0.20
No. of mouseclicks	0.42	0.25	0.30		0.21	-0.21
Remark count	0.35	**0.33**		-0.29		**0.79**
Help count			**-0.67**	-0.42	**0.58**	
Error message count		0.31	-0.22	**0.81**	**0.32**	0.21

Model Results: Item and Paradata terms

Figure 2.6 plots the log-odds of a flag for the item model, paradata model, and 'item + paradata' model; the scaling factors for these models were 0.93, 0.92, and 0.92 respectively. We display all item-level results in terms of log-odds; some models have interactions for which the display of odds ratios can be misleading. A detailed table containing the coefficient estimates, standard errors and p-values pertaining to this figure is in Appendix 2.D.

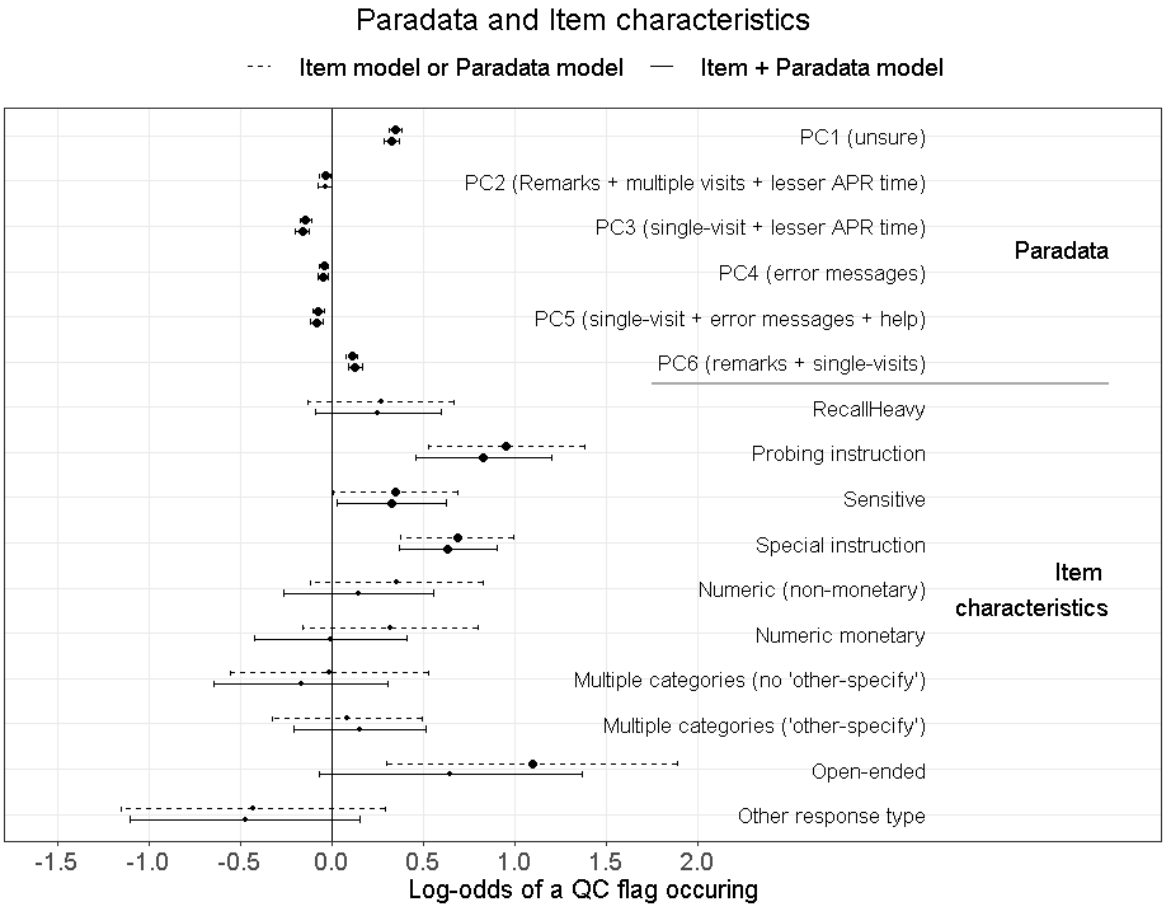

Figure 2.6: Item-level analysis: 'Item + Paradata' model estimates. Log-odds of a QC flag based on the paradata model are represented by dashed lines and those from the 'Item + Paradata' model are shown by solid lines. The horizontal bars are B-H adjusted 90% confidence intervals. The intercept has not been shown in these plots. Point estimates whose confidence intervals do not contain zero are in bold. The grey line to the right of the plot separates the paradata terms from the item characteristic terms.

All 6 PC terms in Figure 2.6 are significant. An increase in the PC1 (unsure interviewing) score by 1 standard deviation increases the odds ratio of a flag by a fairly substantial 41% while the increase in odds ratio associated with PC6 (single-visit remarks) is 12%. The other 4 PCs are associated with a lower odds ratio of a flag, with odds ratios ranging from 0.87 to 0.97. When the item-characteristic terms are added, the effect sizes are approximately the same; PC2 turns insignificant but this term was borderline significant earlier (p-value = 0.088).

Turning to the item model, we do not find evidence of a relationship for the RecallHeavy variable. However, items that have a probing instruction are more likely on average to have a flag occurring, compared to items that do not have such an instruction (odds ratio = 2.6). Sensitive items and items with instructions are also associated with higher odds ratio of a flag (1.42 and 1.99 respectively). Among the response type terms, only the effect for the open-ended response type is significant; compared to a binary response item, on average an open-ended response item has a much greater odds ratio of a flag occurring (odds ratio = 3). The CI for this effect is large given that we have only 4 open-ended items in the analysis; in general, the CIs of the item characteristic terms are much wider than those of the PC terms. In the 'Item + Paradata' model, the effect for the open-ended response type becomes insignificant with a substantial reduction in effect size (odds ratio reduces from 3.00 to 1.92). The odds ratio for items having a probing instruction also reduces from 2.61 to 2.29. These results imply that paradata contain item characteristic effects associated with interviewing quality.

We further pursue this line of analysis by interacting the PC terms with item characteristics, results of which are shown in Figure 2.7; data pertaining to this figure is in Appendix 2.F. PC4 ('error messages') is the only PC that does not have an associated interaction and its main effect too is no longer significant. The remaining 5 of the 6 PCs have significant interactions with the item characteristics and different PCs interact with different characteristics. The PC explaining most of the variance in the data - PC1 (unsure interviewing) - interacts with probing, sensitive, and special instruction characteristics, PC2 interacts with response type, PC3 and PC5 interact with recall-heavy characteristic, and PC6 interacts with the special instruction characteristic. The presence of the recall-heavy characteristic in 2 PCs is interesting; while it itself does not have an overall effect (as seen in Figure 2.6), it plays a moderating role between paradata and interviewing quality. Effect sizes are more or less unchanged in the full model (when the non-paradata terms are added), except for some change in the open-ended main effect. The interaction terms involving PC5 and PC6 are seen to be non-significant in the full model.

Given the interactions in Figure 2.7, to facilitate interpretations we predict the probability of a flag at 3 PC score levels - average, 2 standard deviations above the average ('high'), and 2 standard deviations below the average ('low') - for different item characteristics, all else being equal. The item response type is assumed to be binomial unless stated otherwise. These predictions ignore the random effects. We select the first 3 PCs to conduct this analysis. We start with the interaction of PC1 with the probing instruction and sensitive item characteristics. Figure 2.8 computes flag predictions when an item has a probing instruction. The magnitudes of the predicted probabilities are small since the overall flag rate is only 4.1%. In general, for a given PC1 score level, items that have a probing instruction have a higher chance of a flag occurring than those that don't

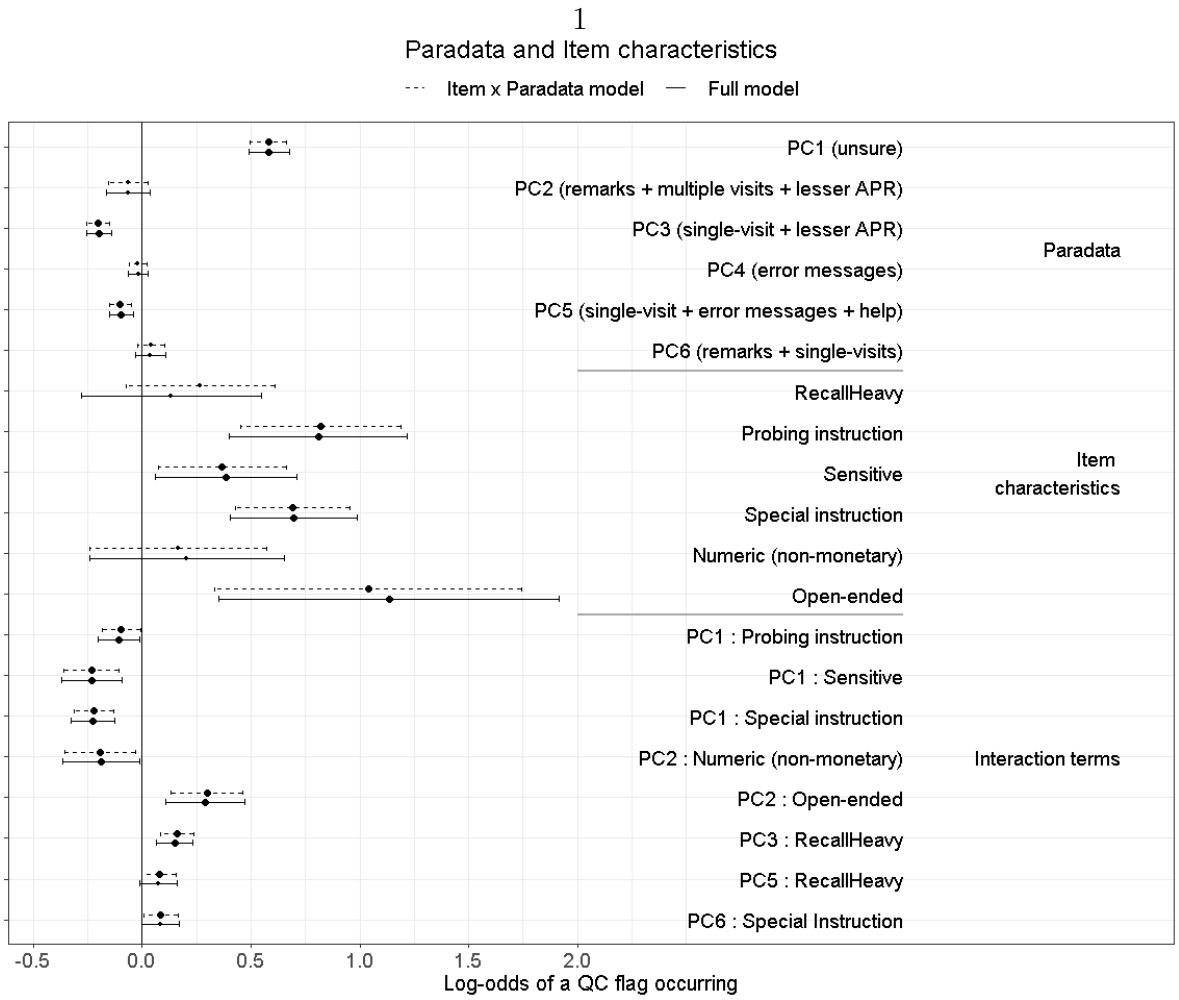

Figure 2.7: Item-level analysis: 'Item x Paradata' and full model estimates. Log-odds of a QC flag based on the 'Item x Paradata' model are represented by dashed lines and those from the full model are shown by solid lines. The horizontal bars are B-H adjusted 90% confidence intervals. The intercept has not been shown in these plots. Odds ratios for response types that are not significant are not shown for cleaner display. Point estimates whose confidence intervals do not contain zero are in bold. The grey lines to the right of the plot separate the paradata terms, the item characteristic terms, and the interactions.

have such an instruction. However, the predicted flag probability for an item that has a probing instruction but a low PC1 score (i.e., straightforward interviewing), is *lower* than the situation for a non-probing instruction item but having a high PC1 score (unsure interviewing).

The scenarios involving sensitive items are shown in Figure 2.9. We find that when the PC1 score is high, the predicted flag probability for a non-sensitive item is *higher* than a sensitive item; it is more understandable to have remarks, more mouseclicks etc. (the prominent measures for PC1) when the item is sensitive than when not.

Turning to PC2, Figure 2.10 plots the predicted flag probabilities for 3 different response types. From the 1st panel, we see that PC2 scores have no impact on the reference binary response type level. The trends for the numeric non-monetary and open-ended

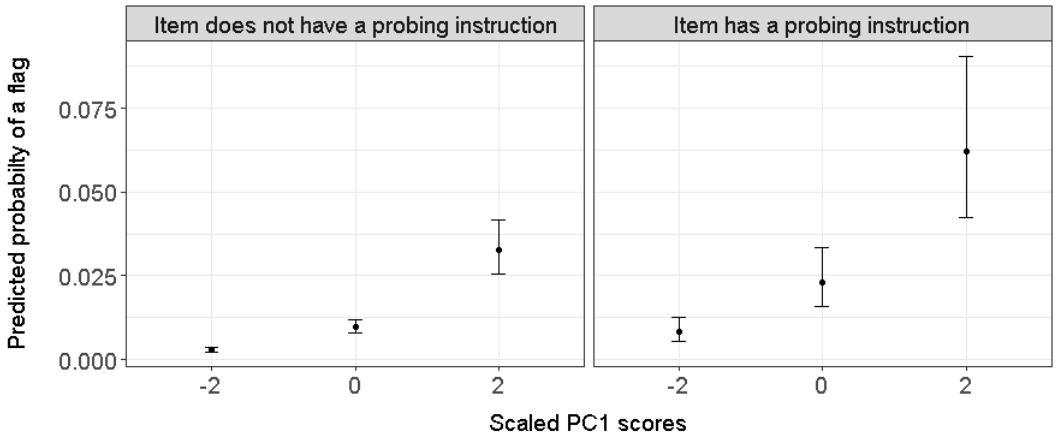

Figure 2.8: Predicted flag probabilities: Interaction between PC1 and the probing instruction indicator. The vertical bars are 90% confidence intervals.

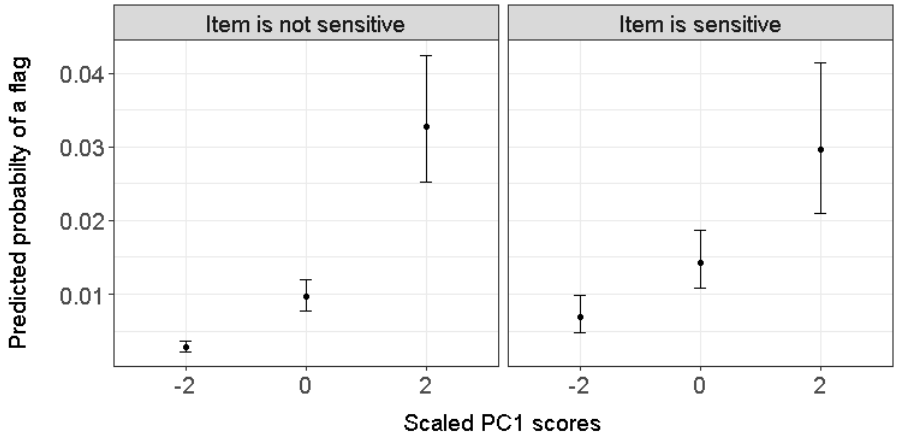

Figure 2.9: Predicted flag probabilities: Interaction between PC1 and the sensitive item indicator. The vertical bars are 90% confidence intervals.

response types go in opposite directions. As the PC2 score increases, the predicted flag probabilities for open-ended items increase; speeding, multiple visits, and remark making are predictive of flag occurrences for this response type. Why does the trend for the numeric non-monetary response type go in the other direction? An example of this response type is item F3 in the PSID questionnaire which asks about time spent on housework in an average week. Consulting the questionnaire objectives manual for this question [3] shows that if the respondent is unable to give an exact number of hours despite probing, interviewers are asked to make a remark. They are also asked to make a remark if roomers or boarders live in the housing unit but the respondent cannot separate time spent on cleaning their rooms (this would be classified as income-producing work). Thus, remark making for this item would actually be an indicator that the interviewer has tried to probe the respondent and is conscientious about remark-making when exact answers are not forthcoming. Further, 871 of the 8789 interviews (9.9%) for this item

[3] `ftp://ftp.isr.umich.edu/pub/src/psid/questionnaires/q2015.pdf`

involved multiple visits. It is not obvious why this should occur but the data do show a correlation between remark-making and revisits for this item. Of the 871 interviews with multiple item visits, 83 also involve a remark for this item (9.5%). In comparison, only 1.8% of the single-visit interviews (149 of 7918 interviews) had a remark. Perhaps good interviewers go back to an item to make a remark when they find discrepancies in the response to this item and a related response to another item.

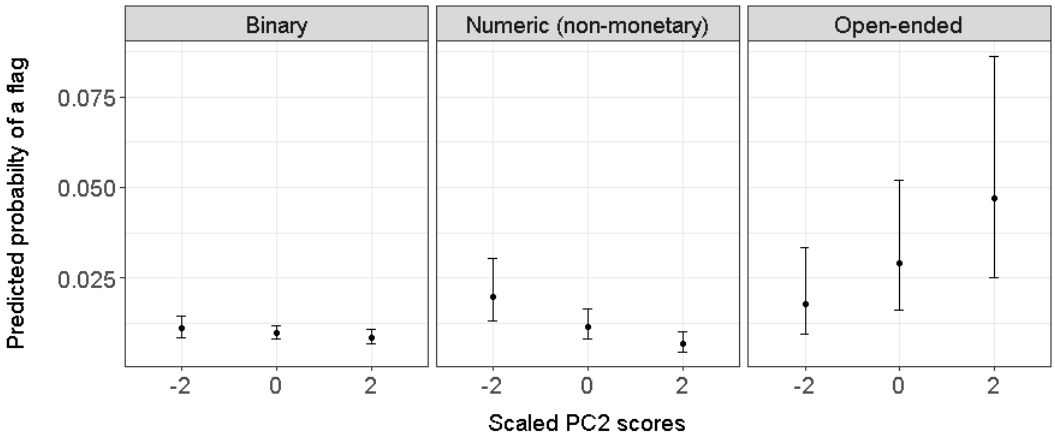

Figure 2.10: Predicted flag probabilities: Interaction between PC2 and 3 item response types. The vertical bars are 90% confidence intervals.

Finally, Figure 2.11 plots the flag probabilities for the recall-heavy item characteristic for PC3. The tendency to not undertake multiple visits and not access help possibly signifies confidence in interviewing which reflects in higher PC scores reducing the chances of a flag occurring. However, this PC is also associated with low APR times. This could be a drawback for recall-heavy items that may require slower questioning and more probing. This may be the reason why the drop in flag probabilities as the PC3 scores increase is not as dramatic as that for items that are not recall-heavy. Here again, we see how item characteristics are moderating item-level paradata effects.

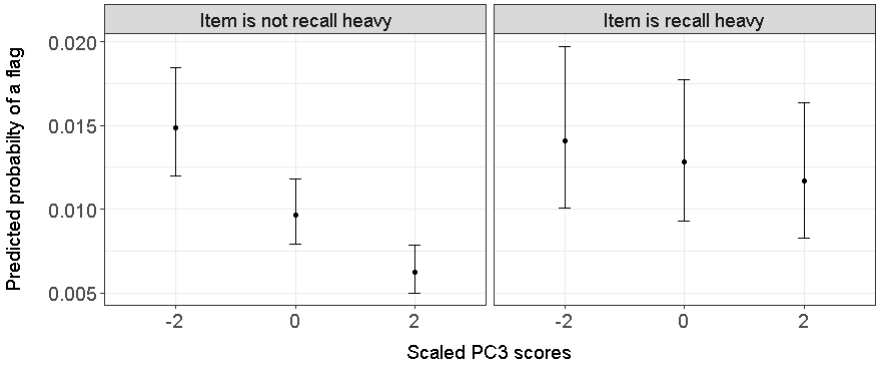

Figure 2.11: Predicted flag probabilities: Interaction between PC3 scores and the recall-heavy indicator. The vertical bars are 90% confidence intervals.

Model Results: Item and Non-paradata terms

We repeat the above analysis for the non-paradata variables. Figure 2.12 shows the log-odds of a flag for the item characteristics model, non-paradata model, and 'item + non-paradata' model. The direction of effects in the non-paradata model are similar to that we saw in the interview-level model. Being a male respondent and having higher levels of education are associated with lower odds ratio of a flag. In contrast, older respondents, the first 2 interviews, male interviewers, and the CV of interviewer daily workload are, on average, associated with higher odds ratio of a flag. When non-paradata terms are added to the item characteristics, there are generally no changes in the effect sizes or significance, the open-ended response type being an exception.

In the 'Item x Non-Paradata' model, we only found the interaction of the recall-heavy characteristic with the 'college graduate and above' interviewer education level (Figure 2.13). For a recall-heavy item, on average, the odds ratio of a flag when the interview is conducted by a college graduate are 37% higher compared to when the interview is done by a high school graduate. When the paradata terms are added to this model, the respondent and interview gender terms are no longer significant, and there are reductions in effect sizes for respondent age and education. The effects for the first two interviews and CV of interviewer workload continue to remain significant with no changes in magnitudes. Appendices 2.E and 2.G contain detailed data pertaining to Figures 2.12 and 2.13.

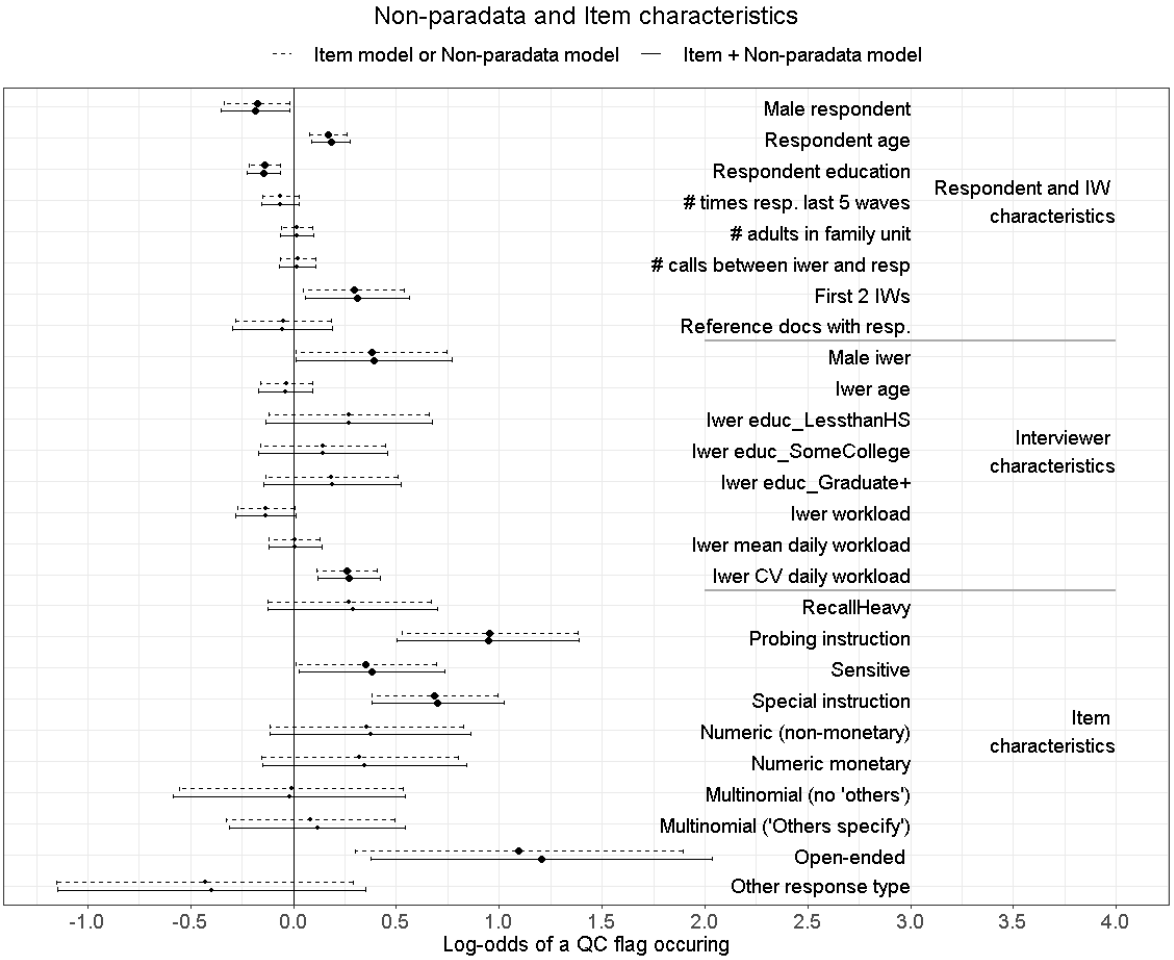

Figure 2.12: Item-level analysis: 'Item + Non-paradata' model estimates. Log-odds of a QC flag based on the only-item or only non-paradata model are shown by dashed lines and those from the 'Item + Non-paradata' model are shown by solid lines. The horizontal bars are B-H adjusted 90% confidence intervals. The intercept has not been shown in these plots. Point estimates whose confidence intervals do not contain zero are in bold. The grey lines to the right of the plot separate the respondent and interview terms, the interviewer terms, and the interactions.

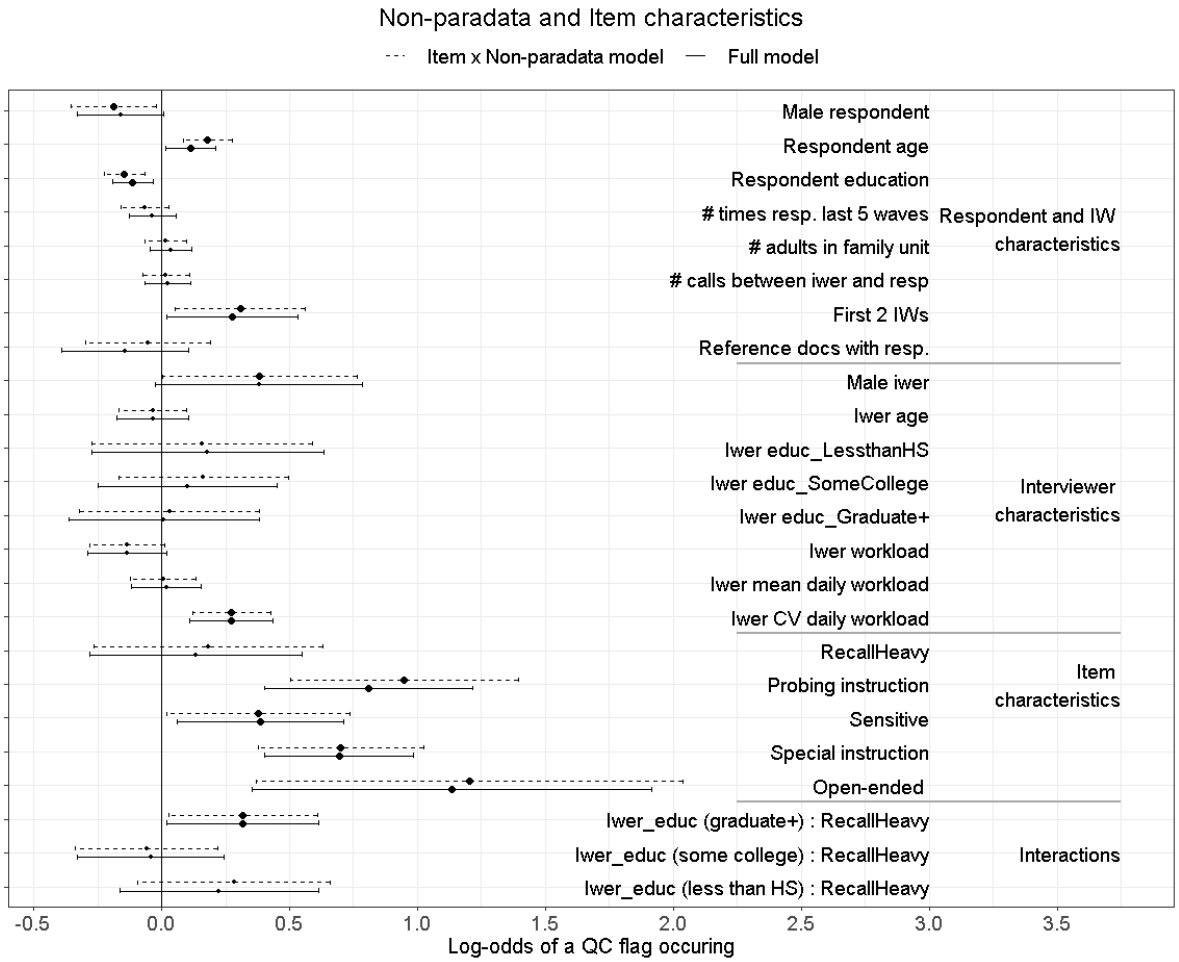

Figure 2.13: Item-level analysis: 'Item x Non-paradata' model estimates. Log-odds of a QC flag based on the item x non-paradata model are shown by dashed lines and those from the full model are shown by solid lines. The horizontal bars are B-H adjusted 90% confidence intervals. The intercept has not been shown in these plots. Point estimates whose confidence intervals do not contain zero are in bold. The grey lines to the right of the plot separate the respondent and interview terms, the interviewer terms, and the interactions. Among the response types, only the significant 'open ended' category is shown.

Model Results: Variance components and model fit

Table 2.8 shows the variance components from our models. We see that in the null model, the random item intercept has the largest variance, followed by that of the interview and interviewer random intercepts. Given the interactions between paradata and item characteristics, the 'Item x Paradata' model explains 63% of the scaled between-item variance compared to 48% explained by the 'Item x Non-paradata' model. However, the explicit inclusion of interviewer-level characteristics in the 'Item x Non-paradata' model allows it to explain 36% of the between-interviewer variance. The between-interviewer variance is seen to increase in the 'Item x Paradata' model which indicates that controlling for phenomena represented by the paradata-item interactions is uncovering the 'true' difference between interviewers. The full model expectedly draws on the relative strengths

of the subset models. However, for a given variance component, we do not see any incremental explanation over the better-performing subset model, e.g., the between-item variance in the full model is almost the same as that in the 'Item x Paradata' model. This indicates that there is a lot of duplication in the sources of variance explained by the 'Item x Paradata' and 'Item x Non-paradata' models.

Table 2.8: Item-level analysis - Model variance components and model fit summary.

	Null	Item x Paradata	Item x Non-paradata	Full
$\hat{\sigma}^2_{item}$	0.839	0.311	0.433	0.319
$\hat{\sigma}^2_{iw}$	0.455	0.385	0.355	0.340
$\hat{\sigma}^2_{iwer}$	0.272	0.292	0.174	0.213
AIC	11830	11243	11732	11225
BIC	11865	11521	12019	11670

In terms of the AIC and BIC criteria, the 'Item x Paradata' model performs the best. We conducted an ROC curve analysis to further assess the performance of these models, the results of which are shown in Figure 2.14. The plot also includes results for the paradata model as a benchmark. Arriving at the optimal probability cut-offs and AUC was done using Youden's method (Youden 1950). In conducting the predictions, we only conditioned on the specific interviewers but not on items and interviews. Conditioning on specific items would also be reasonable and the AUCs would be higher. However, we wanted to check how the models could perform if only the item characteristics were included and not the specific items themselves.

Figure 2.14 shows that the paradata model, with only 6 inputs, is able to correctly predict that a case has a flag for 68% of the actually flagged cases (true positive rate) and incorrectly predicts a flag for 28% of the cases for which there was actually no flag (false positive rate). The AUC for this model is 0.77. In comparison, the 'Item x non-paradata' model despite having 29 terms (excluding the intercept) has a lower true positive rate at 63% with only a 1 percent point lower false positive rate than the paradata model. This establishes the efficiency of the paradata model.

When item characteristics interact with the paradata inputs (paradata x item model), we make gains over the paradata model by reducing the false positive rate and also marginally improving the true positive rate. Finally, the full model increases the true positive rate over the 'Item x Paradata' model by 4% but also increases the false positive rate. A small true flag rate (as in PSID's case) means that an overwhelming number of cases have no flags. This means that even small increases in the false positive rate can translate into large inefficiencies. Therefore, unless there is a strong need to only focus

on the true positive rate, the 'Item x Paradata' model should be preferred.

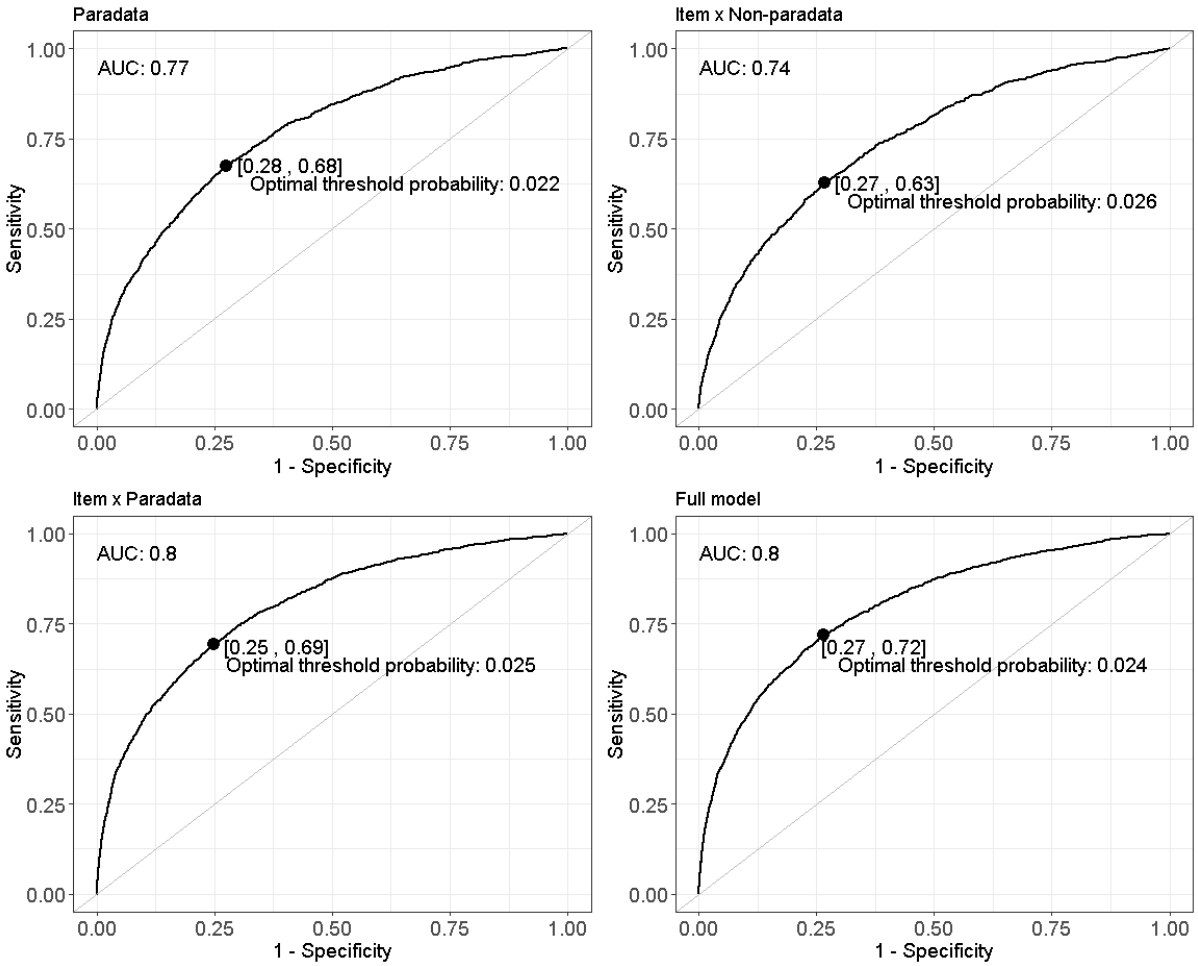

Figure 2.14: ROC analyses for the item-level models. Each panel corresponds to one model and contains information on the Area Under the Curve (AUC), the false positive rate (1 - specificity; horizontal axis), the true positive rate (sensitivity; vertical axis) and the optimal cut-off probability, computed using Youden's method.

Figure 2.15 shows the diagnostic plots for the item-level model; we see no apparent issues.

2.6 Discussion

The results of this study show fairly strong evidence of associations between paradata patterns and interviewing quality. A critical lesson is that a multivariate approach to using paradata is necessary for quality control; it is insufficient, for example, to only focus on 'speeding'. Adopting two analysis levels, at the interview-level and item-level item, aided our understanding of paradata, e.g., the aspect of speeding was clearer through the interview-level analysis. On the other hand, the item-level analysis showed how item characteristics moderate associations between item-level paradata and interviewing quality.

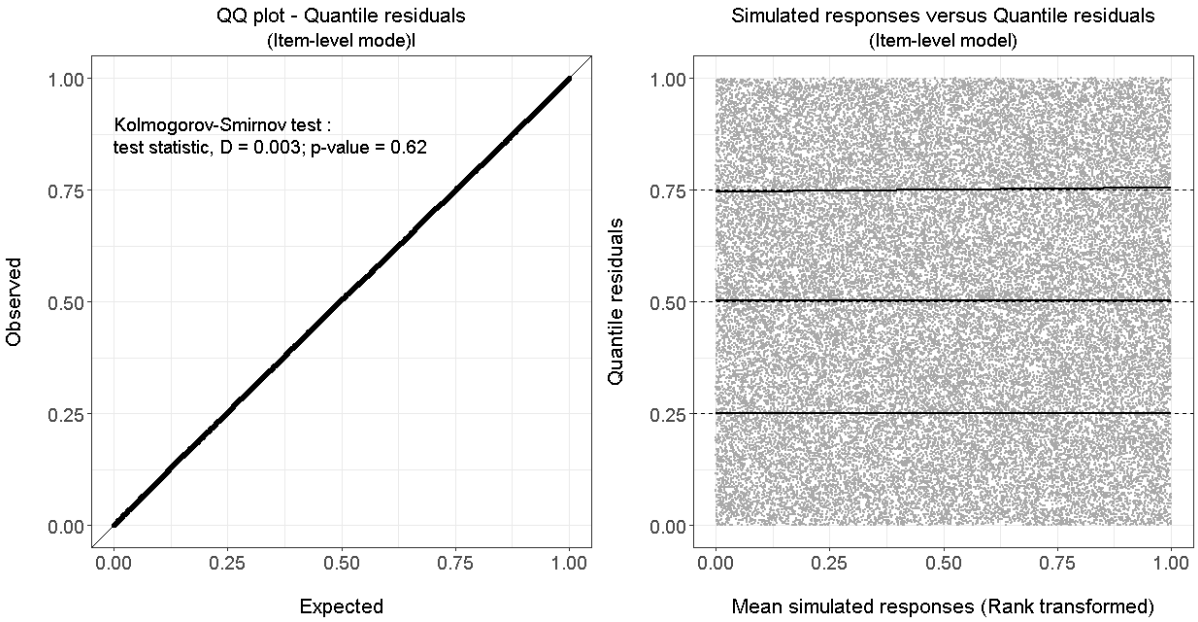

Figure 2.15: Item-level model diagnostics. The left panel compares the quantile residuals to draws from a uniform distribution. Each point in the right panel is the mean simulated response (across 1000 simulations) for an observation in the data (there are therefore 44927 points in this panel). The solid lines correspond to the quantile regression lines. These lines match the dotted benchmark lines.

Many of the respondent effects on interviewing quality disappeared or were reduced when we added the paradata terms into the model, indicating that paradata are capturing interviewing challenges arising due to specific respondent characteristics. One possible mechanism is as follows: older respondents may face more recall issues (Knauper 1999) and therefore require more probing. This would usually take more time. Impatient interviewers who speed-through in such situations would get flagged on account of 'lack of probing'. Such cases would also be associated with a shorter APR time. The evidence suggests that paradata are not only able to capture information present in the non-paradata terms, but are also able to pick up nuances in interviewer behavior such as probing which may not be fully accounted for by only interviewer demographics. This is why the paradata models did better than the non-paradata models on the basis of AIC, BIC, and predictive performance - and far more efficiently (fewer terms).

Interviewers are typically aware when interviews are recorded since consent is obtained from the respondent only in those situations. McGonagle et al. (2015) show that in such situations, interviewers adjust their interviewing pace and some measures of data quality improve. This means that the effects we found in our study could actually be larger in the wider population of interviews and interviewers.

There were limitations in our research: First, we did not have inter-rater reliability scores to evaluate the QC coding. Second, our analysis on predictive power could be overly optimistic since we are predicting the outcome for the same set of data that we

built the model on. For this research, we wanted to utilize all available information since our goal was primarily explanatory and therefore did not split the data into training and test sets. But a pilot project using the paradata models to predict flags in the 2013 PSID data yielded good results. Third, our interview-level analysis ignores item heterogeneity that could lead to inaccurate inferences. Fourth, our results are conditional on the measures we defined; perhaps other more creatively defined measures could yield better results. Fifth, our analysis could have benefited from richer item-level data such as that used in Couper and Kreuter (2013) or Olson and Smyth (2015). At the interviewer-level, one variable which is potentially important is that of interviewer experience (Bailar et al. 1977; Lipps 2007; Couper and Kreuter 2013; West and Blom 2016), which we were unable to access. Sixth, our research cannot confirm the causal mechanisms that are generating our effects, e.g., despite our reasoning, we cannot confirm why making remarks are generally an indication of inadequate interviewing. Investigating these causal mechanisms using methods such as more detailed behavior coding is another important step to put our understanding on a stronger footing. Finally, we undertook an analysis to judge differences between the evaluated and non-evaluated interviews with respect to respondent and interview characteristics; evaluated respondents were older, associated with lesser calls, and interviews occurred earlier in interview sequences. The comparisons are shown in Table 2.9. The p-values based on the Mann-Whitney-Wilcoxon rank sum test indicate differences in the distributions for five of the eight variables (using a significance level of 0.05). These findings might affect the generalizability of the findings.

	Evaluated interviews (555 interviews)						Non-evaluated interviews (8411 respondents)						p-value (diff. in prop.)
Male respondents:	38%						40%						0.46
Respondents with ref. documents available:	14%						17%						0.11
	Min.	Q1	Median	Mean	Q3	Max.	Min.	Q1	Median	Mean	Q3	Max.	p-value (M-W-W)
Respondent education *(years)*	0	12	13	13.5	16	17	0	12	14	13.7	16	17	0.08
Respondent age *(years)*	20	33	48	48	60	94	18	32	42	45	57	97	< 0.001
# waves as respondent in the last 5 waves	1	4	5	4.2	5	5	1	3	5	4.1	5	5	0.001
# adults in the respondent's family unit	1	1	2	1.7	2	5	1	1	2	1.8	2	7	0.02
# calls between respondent and interviewer	1	3	5	14.4	15	139	1	4	8	20.8	23	180	< 0.001
IW sequence in IWER's workload	1	11	37	45	72	174	1	26	53	59.3	86	211	< 0.001

Table 2.9: Comparison of respondent characteristics: Evaluated vs Non-evaluated interviews. Data excludes respondents that are proxy and institutionalized. p-values for characteristics 3-8 are based on the Mann-Whitney-Wilcoxon (M-W-W) rank sum test.

Even surveys with a well-established Computer Audio-Recorded Interviewing (CARI) infrastructure can afford to listen to only a limited number of interviews. However, interest is in the population of interviews. The steps outlined in this research can be used to build models that use paradata to predict interviewing quality for all interviews. For studies such as PSID, models could be built on one wave of data and QC flags predicted

for incoming interviews of the fresh wave based on their paradata patterns. One could have a two-level QC process: one that isolates interviews with a large predicted proportion of errors (interview-level model) and another that focuses on specific items across interviews (item-level model). Future research can experiment with other predictive modeling approaches beginning with relatively simpler steps such as introducing interactions between the paradata and non-paradata terms. Analysts can also consider the following suggestions: First, interview-level PC terms can be added to the item-level model to provide more context, e.g., relatively shorter APR times for an item in a relatively longer interview could probably be subject to a higher chances of a flag; Second, incorporate item-specific slopes in the model in equation 2.4; and third, experiment with other paradata measures. Some measures worth considering are: splitting keycounts between those spent on the response and those on remarks; extracting information on whether a mouseclick was actually used to change a response or if it was simply an idle click; extracting information on the use of backspaces by the interviewer; explicit coding of the return key; incorporating information on whether an item revisit is due to an error message; incorporating additional measures such as 'access to help in multiple item visits'.

In sum, we regard this research as the first step in understanding the interplay between paradata, interviewing quality, and interviewer effects. An important next step would be to replicate this research on different CATI and CAPI surveys. Not only will this strengthen the evidence on hand, but if we find common associations of paradata patterns with interviewing quality across surveys, inferences could potentially be transported to surveys where it might be too expensive to undertake recordings and conduct behavior coding.

Appendix

2.A List of items and their covariates included in the item-level analysis.

Table 2.A.1: List of the 170 items used in the item-level analysis and their characteristics.

No.	Item	Probing instruc.	Special instruc.	Sensitive	Recall heavy	Response type
1	Section_A.A19	1	0	0	0	Multinomial (no 'others')
2	Section_A.A20	0	0	1	1	Numeric monetary
3	Section_A.A21	0	0	1	0	Numeric monetary
4	Section_A.A22	0	0	1	0	Numeric monetary
5	Section_A.A23	0	0	0	0	Binary
6	Section_A.A28	0	0	0	1	Binary
7	Section_A.A3	0	0	0	1	Multinomial (no 'others')
8	Section_A.A31	1	1	0	0	Numeric monetary
9	Section_A.a31b	0	0	0	0	Binary
10	Section_A.a31c	0	0	0	0	Binary
11	Section_A.A31Per	0	0	0	0	Multinomial (others spec.)
12	Section_A.A32	0	0	0	0	Binary
13	Section_A.A33	0	1	0	0	Binary
14	Section_A.A37A	0	0	0	1	Binary
15	Section_A.A4	1	0	0	0	Multinomial (others spec.)
16	Section_A.A40_[1]	1	0	0	0	Multinomial (others spec.)
17	Section_A.A42	0	0	1	0	Multinomial (no 'others')
18	Section_A.A42A	0	1	1	0	Numeric monetary
19	Section_A.A42APer	0	0	0	0	Multinomial (others spec.)
20	Section_A.A42B	0	1	1	0	Numeric monetary
21	Section_A.A42BPER	0	0	0	0	Multinomial (others spec.)
22	Section_A.A42C	0	1	1	0	Numeric monetary
23	Section_A.A43	0	1	1	0	Numeric monetary
24	Section_A.A43Per	0	0	0	0	Multinomial (others spec.)
25	Section_A.A44	0	1	1	0	Numeric monetary
26	Section_A.A44Per	0	0	0	0	Multinomial (others spec.)
27	Section_A.A45	0	0	1	0	Binary
28	Section_A.A46	0	0	0	0	Binary
29	Section_A.A57[1].A57a	0	1	0	0	Binary
30	Section_A.A57[1].A57b	0	0	0	0	Multinomial (no 'others')
31	Section_A.A57[1].A57e	0	0	0	0	Binary
32	Section_A.A57[1].A57l	0	0	0	0	Multinomial (no 'others')
33	Section_A.A6A	0	0	0	0	Binary
34	Section_A.A8	0	1	0	0	Numeric non-monetary

No.	Item	Probing instruc.	Special instruc.	Sensitive	Recall heavy	Response type
35	Section_A.MGTE[1].A23a	0	0	1	0	Multinomial (others spec.)
36	Section_A.MGTE[1].A23b	0	0	0	0	Binary
37	Section_A.MGTE[1].A24	0	0	1	1	Numeric monetary
38	Section_A.MGTE[1].A25	0	0	0	1	Numeric monetary
39	Section_A.MGTE[1].A25a1	0	0	1	0	Binary
40	Section_A.MGTE[1].A25a2	0	0	1	0	Binary
41	Section_A.MGTE[1].A25a3	0	0	1	0	Binary
42	Section_A.MGTE[1].A25a4	0	1	1	0	Numeric monetary
43	Section_A.MGTE[1].A25b	0	0	1	0	Numeric monetary
44	Section_A.MGTE[1].A26	0	0	0	0	Numeric non-monetary
45	Section_A.MGTE[1].A27	0	0	1	1	Numeric non-monetary
46	Section_A.MGTE[1].A27A	0	0	0	1	Binary
47	Section_A.MGTE[1].A27F	0	0	0	1	Binary
48	Section_A.MGTE[1].A27G	0	0	0	1	Multinomial (no ‘others’)
49	EmployEHC1.BCDE1[1]	1	0	0	0	Multinomial (others spec.)
50	EmployEHC1.BCDE2	0	0	1	0	Numeric non-monetary
51	EmployEHC1.BCDE3	0	0	0	0	Binary
52	EmployEHC1.BCDE3A	0	0	0	0	Binary
53	Section_BC.BCJobs[1].BC19b	0	0	0	0	Binary
54	Section_BC.BCJobs[1].BC20	1	1	0	0	Open-ended
55	Section_BC.BCJobs[1].BC21	1	1	0	0	Open-ended
56	Section_BC.BCJobs[1].BC21a	0	0	0	0	Open-ended
57	Section_DE.DEJobs[1].DE19b	0	0	0	0	Binary
58	Section_F.F11	0	1	0	0	Binary
59	Section_F.F14	0	1	0	0	Binary
60	Section_F.F18F22	1	0	1	0	Numeric monetary
61	Section_F.F18F22Per	0	0	0	0	Multinomial (others spec.)
62	Section_F.F19F23	0	0	0	0	Binary
63	Section_F.F21F25	1	0	1	0	Numeric monetary
64	Section_F.F21F25Per	0	0	0	0	Multinomial (others spec.)
65	Section_F.F3	0	1	0	0	Numeric non-monetary
66	Section_F.F47	0	1	0	0	Binary
67	Section_F.F48	0	1	0	0	Numeric non-monetary
68	Section_F.F49Series.F49[1].F49a	0	0	1	0	Numeric non-monetary
69	Section_F.F49Series.F49[1].F49b	1	0	1	0	Others
70	Section_F.F49Series.F49[1].F49b2	1	0	1	0	Multinomial (others spec.)
71	Section_F.F49Series.F49[2].F49a	0	0	1	0	Numeric non-monetary
72	Section_F.F49Series.F49[2].F49b	1	0	1	0	Others
73	Section_F.F49Series.F49[2].F49b2	1	0	1	0	Multinomial (others spec.)
74	Section_F.F77	0	1	1	0	Numeric monetary
75	Section_F.F77Per	0	0	0	0	Multinomial (others spec.)
76	Section_F.F79	0	1	1	0	Numeric monetary
77	Section_F.F8	0	1	0	0	Binary
78	Section_F.F80b	0	1	1	0	Numeric monetary
79	Section_F.F80c	0	1	1	0	Numeric monetary
80	Section_F.F81a	0	1	1	0	Numeric monetary
81	Section_F.F81b	0	1	1	0	Numeric monetary
82	Section_F.F81c	0	1	1	0	Numeric monetary
83	Section_F.F82	0	0	0	0	Binary
84	Section_F.F83	0	0	1	0	Numeric monetary
85	Section_F.F84	0	0	0	0	Binary
86	Section_F.F87	0	0	1	0	Numeric monetary
87	Section_F.F87Per	0	0	0	0	Multinomial (others spec.)
88	Section_F.F88	0	0	1	0	Numeric monetary
89	Section_F.F88Per	0	0	0	0	Multinomial (others spec.)
90	Section_F.F89	0	0	1	0	Numeric monetary
91	Section_F.F89Per	0	0	0	0	Multinomial (others spec.)

No.	Item	Probing instruc.	Special instruc.	Sensitive	Recall heavy	Response type
92	Section_F.F90	0	0	1	0	Numeric monetary
93	Section_F.F90Per	0	0	0	0	Multinomial (others spec.)
94	Section_F.F91	0	0	1	0	Numeric monetary
95	Section_F.Food.FOOD1	0	0	0	1	Multinomial (no 'others')
96	Section_F.Food.FOOD2	0	0	0	1	Multinomial (no 'others')
97	Section_F.Food.FOOD3	0	0	0	1	Multinomial (no 'others')
98	Section_F.Vehicle[1].F53	0	0	0	0	Multinomial (others spec.)
99	Section_F.Vehicle[1].F55	0	0	0	0	Numeric non-monetary
100	Section_F.Vehicle[1].F57	0	0	0	0	Binary
101	Section_F.Vehicle[1].F58	0	1	0	0	Binary
102	Section_F.Vehicle[1].F61	0	0	1	0	Numeric monetary
103	Section_F.Vehicle[1].F64	0	1	1	1	Numeric monetary
104	Section_F.Vehicle[1].F65	0	0	0	0	Binary
105	Section_F.Vehicle[1].F66	0	0	1	1	Numeric monetary
106	Section_F.Vehicle[1].F67	0	0	1	0	Numeric monetary
107	Section_F.Vehicle[1].F67Per	0	0	0	0	Multinomial (others spec.)
108	Section_F.Vehicle[1].F69	0	0	0	0	Numeric non-monetary
109	Section_F.Vehicle[1].F70	0	0	0	0	Numeric non-monetary
110	Section_F.Vehicle[2].F53	0	0	0	0	Multinomial (others spec.)
111	Section_F.Vehicle[2].F55	0	0	0	0	Numeric non-monetary
112	Section_F.Vehicle[2].F57	0	0	0	0	Binary
113	Section_G.G1	0	0	0	0	Others
114	Section_G.G102	0	0	1	0	Binary
115	Section_G.G103	0	0	0	0	Binary
116	Section_G.G12	0	1	0	0	Binary
117	Section_G.G13	0	0	1	1	Numeric monetary
118	Section_G.G14	0	0	0	0	Binary
119	Section_G.G16	0	0	0	0	Binary
120	Section_G.G17f	0	1	0	0	Binary
121	Section_G.G18a	0	0	0	0	Binary
122	Section_G.G25a	0	0	0	0	Binary
123	Section_G.G25b	0	0	0	0	Binary
124	Section_G.G25c	0	0	0	0	Binary
125	Section_G.G25d	0	0	0	0	Binary
126	Section_G.G25e	0	0	0	0	Binary
127	Section_G.G25f	0	1	0	0	Binary
128	Section_G.G25g	0	0	0	0	Binary
129	Section_G.G31	0	1	0	0	Binary
130	Section_G.G37a_[1]	1	0	0	0	Multinomial (others spec.)
131	Section_G.G40_[1]	1	0	0	0	Multinomial (others spec.)
132	Section_G.G44a	0	0	0	0	Binary
133	Section_G.G44b	0	0	0	0	Binary
134	Section_G.G44c	0	0	0	0	Binary
135	Section_G.G44d	0	0	0	0	Binary
136	Section_G.G44e	0	0	0	0	Binary
137	Section_G.G44f	0	0	0	0	Binary
138	Section_G.G44g	0	0	0	0	Binary
139	Section_G.G5	0	0	0	0	Binary
140	Section_G.G99	0	0	0	1	Binary
141	Section_H.H61d2	0	0	0	0	Binary
142	Section_H.H61d3[1]	1	0	0	0	Others
143	Section_H.H61d4	0	0	0	0	Binary
144	Section_H.H61e_H61g[1].H61e_[1]	1	1	0	0	Multinomial (others spec.)
145	Section_H.H61e_H61g[1].H61f_[1]	1	1	0	0	Multinomial (others spec.)
146	Section_H.H61e_H61g[2].H61e_[1]	1	1	0	0	Multinomial (others spec.)
147	Section_H.H61J	0	0	1	0	Numeric monetary
148	Section_H.H61JPer	0	0	0	0	Multinomial (others spec.)

No.	Item	Probing instruc.	Special instruc.	Sensitive	Recall heavy	Response type
149	Section_H.H61k	0	0	0	1	Binary
150	Section_H.H61l[1]	1	0	0	0	Others
151	Section_H.H61m_H61n[1].H61m	0	0	1	0	Numeric non-monetary
152	Section_H.H61m_H61n[1].H61n	0	0	1	0	Numeric non-monetary
153	Section_KL.KL[2].KL74	0	0	0	0	Multinomial (no 'others')
154	Section_KL.KL[2].KL74a	0	1	0	0	Binary
155	Section_KL.KL[2].KL74b	1	0	0	0	Open-ended
156	Section_KL.KL[2].KL84	0	0	0	0	Binary
157	Section_M.M1	0	0	0	1	Binary
158	Section_M.M10	0	0	0	1	Binary
159	Section_M.M11	0	0	0	1	Binary
160	Section_M.M12	0	0	0	0	Binary
161	Section_M.M2	0	0	0	1	Binary
162	Section_M.M2a	0	0	1	1	Numeric monetary
163	Section_M.M3	0	0	0	1	Binary
164	Section_M.M4	0	0	0	1	Binary
165	Section_M.M5	0	0	0	1	Binary
166	Section_M.M6	0	0	0	1	Binary
167	Section_M.M7	0	0	0	1	Binary
168	Section_M.M8	0	0	0	1	Binary
169	Section_M.M9	0	0	0	1	Binary
170	Section_M.MIntro	0	0	0	0	Others

2.B R code for model diagnostics

```
require(DHARMa)
require(qrnn)
require(ggplot2)

#simulate observations
gof_model <- simulateResiduals(model, #name of the model
                               #1000 responses simulated
                               n = 1000,
                               refit = F,
                               #condition on all random effects
                               re.form = NULL)

#mean of the simulated responses for each observation
mean_simresponse <- gof_model$fittedPredictedResponse

#rank transform the mean simulated responses for better visualization
mean_simresponse <- rank(mean_simresponse, ties.method = "average")
mean_simresponse <- mean_simresponse/max(mean_simresponse)

#extract quantile residuals
scaled_resids <- gof_model$scaledResiduals

#data frame for plots
quantresids_data <- data.frame(scaled_resids = scaled_resids, #quantile
                               Expected = runif(gof_model$nObs),
                               mean_simresponse = mean_simresponse)

### Quantile regression
#penalty factor kept as 1 to reduce overfitting
#25th percentile
fit25_nl <- qrnn.fit(x = as.matrix(quantresids_data$mean_simresponse),
                     y = as.matrix(quantresids_data$scaled_resids),
                     n.hidden = 4, iter.max = 1000,
                     n.trials = 1, penalty = 1,
                     tau = 0.25)
quantresids_data$fit25_nl <- qrnn.predict(
  as.matrix(sort(quantresids_data$mean_simresponse)), fit25_nl)

#median
fit50_nl <- qrnn.fit(x = as.matrix(quantresids_data$mean_simresponse),
                     y = as.matrix(quantresids_data$scaled_resids),
                     n.hidden = 4, iter.max = 1000,
                     n.trials = 1, penalty = 1,
                     tau = 0.5)
```

```
quantresids_data$fit50_nl <- qrnn.predict(
  as.matrix(sort(quantresids_data$mean_simresponse)), fit50_nl)

#75th percentile
fit75_nl <- qrnn.fit(x = as.matrix(quantresids_data$mean_simresponse),
                     y = as.matrix(quantresids_data$scaled_resids),
                     n.hidden = 4, iter.max = 1000,
                     n.trials = 1, penalty = 1,
                     tau = 0.75)
quantresids_data$fit75_nl <- qrnn.predict(
  as.matrix(sort(quantresids_data$mean_simresponse)), fit75_nl)

#Kolmogorov-Smirnov test - uniform reference distribution
#used for annotation in the QQ plot
ks.test(quantresids_data$scaled_resids, 'punif')


#QQ plot (theme elements and annotations not shown for brevity)
p_quantresids <- ggplot(quantresids_data,
                        aes(x = sort(Expected),
                            y = sort(scaled_resids))) +
  geom_abline(slope = 1, intercept = 0) +
  ggtitle("QQ plot - Quantile residuals",
  subtitle = "(Interview-level mode)1") +
  xlab("Expected") + ylab("Observed")

#Plot of mean simulated responses against quantile residuals
#(theme elements and annotations not shown for brevity)
p_fitted_quantresids <- ggplot(quantresids_data,
                               aes(x = mean_simresponse,
                               y = scaled_resids)) +

  #quantile regression lines
  geom_line(aes(x = sort(mean_simresponse), y = fit25_nl), size = 1) +
  geom_line(aes(x = sort(mean_simresponse), y = fit50_nl), size = 1) +
  geom_line(aes(x = sort(mean_simresponse), y = fit75_nl), size = 1) +

  #reference lines
  geom_abline(slope = 0, intercept = 0.25, linetype = 2) +
  geom_abline(slope = 0, intercept = 0.50, linetype = 2) +
  geom_abline(slope = 0, intercept = 0.75, linetype = 2) +

  ggtitle("Simulated responses versus Quantile residuals",
  subtitle = "(Interview-level model)") +
  xlab("Mean simulated responses (Rank transformed)") +
  ylab("Quantile residuals")
```

2.C Interview-level model results

Table 2.C.1: Data used to plot the interview-level model plots. These correspond to Figures 2.2 and 2.3. The p-values are Benjamini-Hochberg adjusted p-values. p-values less than 0.1 are in bold. OR stands for 'Odds ratio'.

	Paradata model			**Full model**		
	OR	SE	p-value	OR	SE	p-value
(Intercept)	0.04	0.002	**< 0.001**	0.04	0.004	**< 0.001**
PC1	1.09	0.04	**0.02**	1.06	0.04	0.19
PC2	1.11	0.04	**0.01**	1.11	0.05	**0.08**
PC3	0.98	0.03	0.64	0.98	0.04	0.73
PC4	1.08	0.04	**0.03**	1.06	0.04	0.19
PC5	0.93	0.03	**0.03**	0.95	0.03	0.20
PC6	1.09	0.04	**0.03**	1.07	0.04	0.19
PC7	0.98	0.03	0.64	0.98	0.03	0.72
PC8	0.91	0.03	**0.02**	0.91	0.03	**0.06**

	Non-paradata model					
	OR	SE	p-value			
(Intercept)	0.04	0.004	**< 0.001**			
Male respondent	0.85	0.06	**0.04**	0.89	0.06	0.19
Respondent age	1.08	0.04	**0.09**	0.98	0.04	0.79
Respondent education	0.92	0.03	**0.03**	0.94	0.03	0.19
# times resp. last 5 waves	0.94	0.03	0.16	0.96	0.03	0.57
# adults in family unit	1.03	0.03	0.51	1.03	0.04	0.66
# calls between IWER and resp	1.03	0.04	0.51	1.03	0.04	0.58
First 2 IWs	1.29	0.12	**0.03**	1.27	0.12	**0.07**
Reference docs with resp.	0.96	0.09	0.79	0.97	0.09	0.85
Male IWER	1.42	0.23	**0.09**	1.27	0.21	0.28
IWER age	1.00	0.06	0.97	0.99	0.06	0.92
IWER educ_LessthanHS	1.15	0.20	0.56	1.11	0.19	0.73
IWER educ_SomeCollege	1.02	0.14	0.94	1.00	0.14	0.97
IWER educ_Graduate+	1.13	0.16	0.53	1.07	0.15	0.78
IWER workload	0.89	0.06	0.11	0.90	0.06	0.19
IWER mean daily workload	0.99	0.05	0.94	0.99	0.05	0.90
IWER CV daily workload	1.26	0.08	**0.003**	1.26	0.08	**0.01**

2.D Item-level results: Item, Paradata and Item + Paradata models

Table 2.D.1: Data used to plot the Item, Paradata, and Item + Paradata model estimates. These correspond to Figure 2.6. The p-values are Benjamini-Hochberg adjusted p-values. p-values less than 0.1 are in bold. For the item response type variable, 'binary' is used as the reference category.

	Paradata model			**Item + Paradata model**		
	Log-odds	SE	p-value	Log-odds	SE	p-value
(Intercept)	-3.85	0.09	**< 0.001**	-4.33	0.12	**< 0.001**
PC1	0.35	0.02	**< 0.001**	0.33	0.02	**< 0.001**
PC2	-0.03	0.02	**0.09**	-0.03	0.02	0.12
PC3	-0.14	0.02	**< 0.001**	-0.16	0.02	**< 0.001**
PC4	-0.04	0.01	**0.004**	-0.04	0.01	**0.01**
PC5	-0.07	0.02	**< 0.001**	-0.08	0.02	**< 0.001**
PC6	0.11	0.02	**< 0.001**	0.13	0.02	**< 0.001**
	Item model					
	Log-odds	SE	p-value			
(Intercept)	-4.39	0.13	**< 0.001**			
RecallHeavy	0.27	0.20	0.24	0.25	0.18	0.20
Probing instruction	0.96	0.21	**< 0.001**	0.83	0.19	**< 0.001**
Sensitive	0.35	0.17	**0.09**	0.33	0.15	**0.06**
Special instruction	0.69	0.15	**< 0.001**	0.64	0.14	**< 0.001**
Numeric item	0.36	0.24	0.24	0.15	0.21	0.53
Numeric non-monetary item	0.32	0.24	0.24	-0.004	0.21	0.99
Multinomial (no 'others') item	-0.01	0.27	0.97	-0.17	0.25	0.53
Multinomial ('Others specify') item	0.08	0.21	0.75	0.16	0.19	0.48
Open-ended item	1.10	0.40	**0.02**	0.65	0.37	0.12
Other response type	-0.43	0.36	0.29	-0.47	0.33	0.20

2.E Item-level results: Item, Non-paradata and Item + Non-paradata models

Table 2.E.1: Data used to plot the Item, Non-paradata, and Item + Non-paradata model estimates. These correspond to Figure 2.12. The p-values are Benjamini-Hochberg adjusted p-values. p-values less than 0.1 are in bold. For the item response type variable, 'binary' is used as the reference category. For interviewer education, 'High school/GED' is used as the reference category.

	Non-paradata model			**Item + Non-paradata model**		
	Log-odds	SE	p-value	Log-odds	SE	p-value
(Intercept)	-3.91	0.13	**< 0.001**	-0.18	0.08	**0.07**
Male respondent	-0.18	0.08	**0.07**	0.18	0.05	**< 0.001**
Respondent age	0.17	0.04	**0.001**	-0.15	0.04	**0.001**
Respondent education	-0.14	0.04	**0.001**	-0.06	0.04	0.26
# times resp. last 5 waves	-0.06	0.04	0.27	0.02	0.04	0.74
# adults in family unit	0.02	0.04	0.72	0.02	0.04	0.74
# calls between iwer and resp	0.02	0.04	0.72	0.31	0.12	0.04
First 2 IWs	0.29	0.12	**0.05**	0.31	0.12	**0.04**
Reference docs with resp.	-0.05	0.11	0.72	-0.05	0.12	0.74
Male iwer	0.38	0.18	**0.08**	-0.04	0.06	0.70
Iwer age	-0.03	0.06	0.72	0.27	0.20	0.27
Iwer educ_LessthanHS	0.27	0.19	0.27	0.14	0.15	0.47
Iwer educ_SomeCollege	0.14	0.15	0.48	0.19	0.16	0.37
Iwer educ_Graduate+	0.19	0.16	0.37	-0.14	0.07	0.13
Iwer workload	-0.13	0.07	0.12	0.01	0.06	0.93
Iwer mean daily workload	0.01	0.06	0.93	0.27	0.07	**0.001**
Iwer CV daily workload	0.26	0.07	**0.001**	0.27	0.07	**0.001**

	Item model					
	Log-odds	SE	p-value	Log-odds	SE	p-value
(Intercept)	-4.39	0.13	**< 0.001**			
RecallHeavy	0.27	0.20	0.24	0.29	0.20	0.26
Probing instruction	0.96	0.21	**< 0.001**	0.95	0.22	**< 0.001**
Sensitive	0.35	0.17	**0.09**	0.38	0.17	**0.08**
Special instruction	0.69	0.15	**< 0.001**	0.70	0.16	**< 0.001**
Numeric item	0.36	0.24	0.24	0.37	0.24	0.25
Numeric non-monetary item	0.32	0.24	0.24	0.35	0.24	0.26
Multinomial (no 'others') item	-0.01	0.27	0.97	-0.02	0.28	0.94
Multinomial ('Others specify') item	0.08	0.21	0.75	0.12	0.21	0.70
Open-ended item	1.10	0.40	**0.02**	1.20	0.41	**0.01**
Other response type	-0.43	0.36	0.29	-0.40	0.37	0.39

2.F Item-level results: Item x Paradata and Full models

Table 2.F.1: Data used to plot the Item x Paradata and full model estimates. These correspond to Figure 2.7. The p-values are Benjamini-Hochberg adjusted p-values. p-values less than 0.1 are in bold. For the item response type variable, 'binary' is used as the reference category.

	Item x Paradata model			**Relevant full model terms**		
	Log-odds	SE	p-value	Log-odds	SE	p-value
(Intercept)	-4.31	0.12	**< 0.001**	-4.43	0.16	**< 0.001**
PC1	0.58	0.04	**< 0.001**	0.58	0.04	**< 0.001**
PC2	-0.06	0.05	0.30	-0.07	0.05	0.32
PC3	-0.20	0.03	**< 0.001**	-0.20	0.03	**< 0.001**
PC4	-0.02	0.02	0.44	-0.02	0.02	0.57
PC5	-0.10	0.03	**< 0.001**	-0.10	0.03	**0.001**
PC6	0.04	0.03	0.34	0.04	0.03	0.43
RecallHeavy	0.27	0.18	0.22	0.13	0.20	0.66
Probing instruction	0.82	0.20	**< 0.001**	0.81	0.20	**< 0.001**
Sensitive	0.37	0.16	**0.04**	0.39	0.16	**0.05**
Special instruction	0.69	0.14	**< 0.001**	0.69	0.14	**< 0.001**
Numeric monetary item	0.17	0.21	0.53	0.20	0.22	0.52
Numeric non-monetary item	-0.03	0.22	0.93	0.06	0.22	0.85
Multinomial (no 'others') item	-0.22	0.25	0.49	-0.21	0.25	0.57
Multinomial ('Others specify') item	0.22	0.19	0.34	0.26	0.19	0.32
Open-ended item	1.04	0.37	**0.01**	1.13	0.38	**0.01**
Other response type	-0.68	0.33	**0.08**	-0.63	0.34	0.13
PC1 : Sensitive	-0.23	0.07	**0.001**	-0.23	0.07	**0.003**
PC1 : Probing_instruc.	-0.10	0.05	**0.08**	-0.11	0.05	**0.07**
PC1 : Special_instruc.	-0.22	0.05	**< 0.001**	-0.23	0.05	**< 0.001**
PC2 : Numeric non-monetary item	-0.19	0.09	**0.05**	-0.19	0.09	**0.07**
PC2 : Numeric monetary item	0.00	0.06	0.98	0.00	0.06	0.96
PC2 : Multinomial (no 'others') item	-0.04	0.09	0.79	-0.03	0.09	0.85
PC2 : Multinomial ('others specify') item	-0.01	0.06	0.91	-0.01	0.06	0.91
PC2 : Open-ended item	0.30	0.09	**0.002**	0.29	0.09	**0.004**
PC2 : Other response type	-0.03	0.13	0.91	-0.01	0.13	0.96
PC3 : RecallHeavy	0.16	0.04	**< 0.001**	0.15	0.04	**< 0.001**
PC5 : RecallHeavy	0.08	0.04	0.10	0.08	0.04	0.14
PC6 : Special_instruc.	0.08	0.04	**0.08**	0.08	0.04	0.11

2.G Item-level results: Item x Non-paradata and Full models

Table 2.G.1: Data used to plot the Item x Non-paradata and full model estimates. These correspond to Figure 2.13. The p-values are Benjamini-Hochberg adjusted p-values. p-values less than 0.1 are in bold. For the item response type variable, 'binary' is used as the reference category. For interviewer education, 'High school/GED' is used as the reference category.

	Item x Non-paradata model			**Relevant full model terms**		
	Log-odds	SE	p-value	Log-odds	SE	p-value
(Intercept)	-4.54	0.16	**< 0.001**	-4.43	0.16	**< 0.001**
Male respondent	-0.19	0.08	**0.07**	-0.16	0.08	0.11
Respondent age	0.18	0.05	**< 0.001**	0.11	0.05	**0.05**
Respondent education	-0.14	0.04	**0.001**	-0.11	0.04	**0.01**
# times resp. last 5 waves	-0.06	0.04	0.27	-0.04	0.05	0.58
# adults in family unit	0.02	0.04	0.77	0.04	0.04	0.52
# calls between iwer and resp	0.02	0.04	0.77	0.02	0.04	0.71
First 2 IWs	0.31	0.12	**0.05**	0.28	0.12	**0.07**
Reference docs with resp.	-0.05	0.12	0.77	-0.14	0.12	0.39
Male iwer	0.38	0.18	**0.10**	0.38	0.20	0.11
Iwer age	-0.03	0.06	0.77	-0.03	0.07	0.74
Iwereduc_LessthanHS	0.16	0.21	0.64	0.18	0.22	0.57
Iwereduc_SomeCollege	0.16	0.16	0.49	0.10	0.17	0.70
Iwereduc_Graduate+	0.03	0.17	0.92	0.01	0.18	0.96
Iwer workload	-0.14	0.07	0.13	-0.14	0.08	0.14
Iwer mean daily workload	0.01	0.06	0.94	0.02	0.07	0.85
Iwer CV daily workload	0.27	0.07	**0.001**	0.27	0.08	**0.002**
RecallHeavy	0.18	0.22	0.60	0.13	0.20	0.66
Probing instruction	0.95	0.22	0.001	0.81	0.20	**< 0.001**
Sensitive	0.38	0.17	0.08	0.39	0.16	0.05
Special instruction	0.70	0.16	**< 0.001**	0.69	0.14	**< 0.001**
Numeric non-monetary item	0.38	0.24	0.25	0.20	0.22	0.52
Numeric monetary item	0.35	0.24	0.27	0.06	0.22	0.85
Multinomial ('Others specify') item	0.11	0.21	0.77	0.26	0.19	0.32
Multinomial (no 'others') item	-0.02	0.28	0.94	-0.21	0.25	0.57
Open-ended item	1.20	0.41	**0.01**	1.13	0.38	**0.01**
Other response type	-0.39	0.37	0.49	-0.63	0.34	0.13
Iwereduc_LessthanHS : RecallHeavy	0.28	0.18	0.25	0.22	0.19	0.39
Iwereduc_SomeCollege : RecallHeavy	-0.06	0.14	0.77	-0.04	0.14	0.85
Iwereduc_Graduate+ : RecallHeavy	0.32	0.14	**0.07**	0.32	0.14	**0.07**

References

Auriat, N. (1993). "My Wife Knows Best": A Comparison of Event Dating Accuracy Between the Wife, the Husband, the Couple, and the Belgium Population Register. *Public Opin. Q.*, 57(2):165.

Austin, P. C. and Merlo, J. (2017). Intermediate and advanced topics in multilevel logistic regression analysis. *Stat. Med.*, 36(20):3257–3277.

Bailar, B., Bailey, L., and Stevens, J. (1977). Measures of Interviewer Bias and Variance. *J. Mark. Res.*, 14(3):337.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.*, 67(1).

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2005). False Discovery RateAdjusted Multiple Confidence Intervals for Selected Parameters. *J. Am. Stat. Assoc.*, 100(469):71–81.

Browne, W., Subramanian, S., Jones, K., and Goldstein, H. (2005). Variance partitioning in multilevel logistic models that exhibit overdispersion. *J. R. Stat. Soc. Ser. A (Statistics Soc.*, 168(3):599–613.

Bumpstead, R. (2001). A Practical Application of Audit Trails. In *IBUC 7th Int. Blaise Users Conf.*, Washington DC, USA.

Cannon, A. (2011). Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Comput. Geosci.*, 37(9):1277–1284.

Cheung, G., Piskorowski, A., Wood, L., and Peng, H. (2016). Using Survey Paradata. In *IBUC 15th Int. Blaise Users Conf.*, Washington DC, USA.

Couper, M. (1998). Measuring survey quality in a CASIC environment. In *Proc. Jt. Stat. Meet. Am. Stat. Assoc. Surv. Res. Methods Sect.*, pages 41–49, Alexandria, VA. American Statistical Association.

Couper, M. (2000). Usability Evaluation of Computer-Assisted Survey Instruments. *Soc. Sci. Comput. Rev.*, 18(4):384–396.

Couper, M., Holland, L., and Groves, R. (1992). Developing systematic procedures for monitoring in a centralized telephone facility. *J. Off. Stat.*, 8(1):63–76.

Couper, M., Horm, J., and Schlegel, J. (1997). Using trace files to evaluate the National Health Interview Survey CAPI Instrument. In *Proc. Surv. Res. Methods Sect. ASA*, pages 825–829, Anaheim, CA.

Couper, M. and Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *J. R. Stat. Soc. Ser. A (Statistics Soc.*, 176(1):271–286.

Devonshire, J. (2013). Adding Business Intelligence to Paradata: The Blaise Audit Trail. In *IBUC 15th Int. Blaise Users Conf.*, Washington DC, USA.

Dunn, P. and Smyth, G. (1996). Randomized Quantile Residuals. *J. Comput. Graph. Stat.*, 5(3):236.

Edwards, B., Schneider, S., and Brick, P. (2008). Visual Elements of Questionnaire Design: Experiments with a CATI Establishment Survey. In *Adv. Teleph. Surv. Methodol.*, pages 276–296. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Faraway, J. (2005). *Linear models in R*. Chapman & Hall/CRC, first edition.

Faraway, J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman and Hall/CRC, 1st edition.

Fowler, F. and Mangione, T. (1990). *Standardized Survey Interviewing: Minimizing Interviewer Related Error*. SAGE Publications, Inc., Thousand Oaks, California.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Groves, R. (1989). *Survey Errors and Survey Costs*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Gu, H., Couper, M., Kirgis, N., Parker, S., and Buageila, S. (2013). Using Audit Trail Data for Interviewer Data Quality Management. In *Am. Assoc. Public Opin. Res.*

Hadi, A. and Ling, R. (1998). Some Cautionary Notes on the Use of Principal Components Regression. *Am. Stat.*, 52(1):15–19.

Hansen, S. and Marvin, T. (2001). Reporting on Item Times and Keystrokes from Blaise Audit Trails. In *IBUC 7th Int. Blaise Users Conf.*, Annapolis, USA.

Hansen, S.-E., Couper, M., and Fuchs, M. (1998). Usability evaluation of the NHIS CAPI instrument. In *Proc. Surv. Res. methods Sect. Am. Stat. Assoc.*, pages 929–933.

Hartig, F. (2018). DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. `https://cran.r-project.org/package=DHARMa`.

Hicks, W., Edwards, B., Tourangeau, K., McBride, B., Harris-Kojetin, L., and Moss, A. (2010). Using Cari Tools To Understand Measurement Error. *Public Opin. Q.*, 74(5):985–1003.

Hox, J. (2010). *Multilevel Analysis: Techniques and Applications*. Routledge, New York, NY, second edition.

Hunt, J. (2016). Using Audit Trail data to move from a Black Box to a Transparent Data

Collection Process. In *IBUC 15th Int. Blaise Users Conf.*, Washington DC, USA.

Johnson, S. The NLopt nonlinear-optimization package.

Johnson, T., Parker, V., and Clements, C. (2001). Detection and Prevention of Data Falsification in Survey Research. *Surv. Res. Newsl. from Surv. Res. Lab.*, 32(3):1–2.

Jolliffe, I. (1986). *Principal Component Analysis*. Springer New York, New York, NY.

Joyal, É. (2016). The Uses of Blaise Audit Trail Files at Statistics Canada. In *IBUC 7th Int. Blaise Users Conf.*, Netherlands.

Knauper, B. (1999). The Impact of Age and Education on Response Order Effects in Attitude Measurement. *Public Opin. Q.*, 63(3):347.

Korner-Nievergelt, F., Roth von Felten, S., Guelat, J., Almasi, B., and Korner-Nievergelt, P. (2015). *Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS, and STAN*. Elsevier.

Kreuter, F. (2013). Improving Surveys with Paradata: Introduction. In Kreuter, F., editor, *Improv. Surv. with Parad.*, pages 1–9. Wiley, Hoboken, New Jersey.

Krosnick, J. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Appl. Cogn. Psychol.*, 5(3):213–236.

Krosnick, J. and Alwin, D. (1987). An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opin. Q.*, 51(2):201.

Lee, J. and Lee, S. (2012). Does it Matter WHO Responded to the Survey? Trends in the U.S. Gender Earnings Gap Revisited. *ILR Rev.*, 65(1):148–160.

Lepkowski, J., Couper, M., Hansen, S., Landers, W., Mcgonagle, K., and Schlegel, J. (1998). CAPI Instrument Evaluation: Behavior Coding, Trace Files, and Usability Methods. In *Proc. Surv. Res. Methods Sect. ASA*, pages 917–922, Dallas,TX.

Lipps, O. (2007). Interviewer and Respondent Survey Quality Effects in a Cati Panel. *Bull. Sociol. Methodol. Méthodologie Sociol.*, 95(1):5–25.

Loosveldt, G. and Beullens, K. (2013a). 'How long will it take?' An analysis of interview length in the fifth round of the European Social Survey. *Surv. Res. Methods*, 7(2):69–78.

Loosveldt, G. and Beullens, K. (2013b). The impact of respondents and interviewers on interview speed in face-to-face interviews. *Soc. Sci. Res.*, 42(6):1422–1430.

McGonagle, K., Brown, C., and Schoeni, R. (2015). The Effects of Respondents' Consent to Be Recorded on Interview Length and Data Quality in a National Panel Study. *Field methods*, 27(4):373–390.

Mockovak, W. and Powers, R. (2008). The Use of Paradata for Evaluating Interviewer Training and Performance. In *Proc. Surv. Res. Methods Sect. ASA*, Denver, CO. American Statistical Association.

Moshinsky, D. and Carter, C. (2013). Tracking Interviewer Performance by Measuring Time Spent on Field. In *IBUC 12th Int. Blaise Users Conf.*, Washington DC, USA.

Nicolaas, G. (2011). ESRC National Centre for Research Methods Review paper Survey Paradata : A review. Technical Report January, ESRC National Centre for Research Methods.

Nix, B. (2014). The Effect of Interview Length on Data Quality in the Consumer Expenditure Interview Survey. In *JSM Proceedings, Gov. Stat. Sect.*, pages 1405–1414, Alexandria, VA. American Statistical Association.

Olson, K. and Smyth, J. (2015). The Effect of CATI Questions, Respondents, and Interviewers on Response Time. *J. Surv. Stat. Methodol.*, 3(3):361–396.

Penne, M. and Snodgrass, J. (2003). Analyzing Audit Trails in the National Survey on Drug Use and Health (NSDUH). In *IBUC 8th Int. Blaise Users Conf.*, Copenhagen, Denmark.

Rabe-Hesketh, S. and Skrondal, A. (2008). *Multilevel and Longitudinal Modeling Using Stata.* Stata Press, second edition.

Raudenbush, S. (2008). Many Small Groups. In *Handb. Multilevel Anal.*, pages 207–236. Springer, New York, NY.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1):77.

Self, S. and Liang, K.-Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. *J. Am. Stat. Assoc.*, 82(398):605–610.

Skowronski, J. and Thompson, C. (1990). Reconstructing the dates of personal events: Gender differences in accuracy. *Appl. Cogn. Psychol.*, 4(5):371–381.

Skrondal, A. and Rabe-Hesketh, S. (2007). Redundant Overdispersion Parameters in Multilevel Models for Categorical Responses. *J. Educ. Behav. Stat.*, 32(4):419–430.

Snijders, T. and Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* Sage Publishers, London, 1st edition.

Steve, K., Burks, A.-T., Lavrakas, P., Brown, K., and Hoover, J. (2007). Monitoring Telephone Interviewer Performance. In Lepkowski, J., Tucker, C., M., B., de Leeuw, E., Japec, L., Lavrakas, P., Link, M., and Sangster, R., editors, *Adv. Teleph. Surv. Methodol.*, pages 401–422. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Team, R. C. (2013). R: A Language and Environment for Statistical Computing.

Tourangeau, R., Rips, L., and Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge University Press, Cambridge.

Tourangeau, R. and Yan, T. (2007). Sensitive questions in surveys. *Psychol. Bull.*, 133(5):859–883.

Turner, G., Sturgis, P., and Martin, D. (2015). Can Response Latencies Be Used to Detect Survey Satisficing on Cognitively Demanding Questions? *J. Surv. Stat. Methodol.*, 3(1):89–108.

Vandenplas, C., Loosveldt, G., Beullens, K., and Denies, K. (2017). Are Interviewer Effects on Interview Speed Related to Interviewer Effects on Straight-Lining Tendency in the European Social Survey? An Interviewer-Related Analysis. *J. Surv. Stat. Methodol.*

Wang, K., Kott, P., and Moore, A. (2013). Asessing the relationship between interviewer effects and NSDUH data quality. Technical report, Report prepared by RTI International for Substance Abuse and Mental Health Services Administration.

West, B. and Blom, A. (2016). Explaining Interviewer Effects: A Research Synthesis. *J. Surv. Stat. Methodol.*

West, B. and Sinibaldi, J. (2013). The Quality of Paradata: A Literature Review. In *Improv. Surv. with Parad.*, pages 339–359. John Wiley & Sons, Inc., Hoboken, New Jersey.

Youden, W. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Ypma, J., Borchers, H., and Eddelbuettel, D. (2014). nloptr - R interface to NLopt.

Chapter 3

Does monitoring interviewing quality imply monitoring interviewer effects?

3.1 Why monitor interviewing quality?

Most professionally-run surveys have systems that monitor interviewers for interviewing quality (Tarnai and Moore 2007). These systems have their roots in the recognition that interviewer behaviors can introduce error into survey estimates. Interviewers differ in the way they ask questions (Rustemeyer 1977; Couper et al. 1992), probe respondent answers (Hyman 1954; Mangione et al. 1992), give them feedback (Marquis 1969; Hildum and Brown 1956; Cannell et al. 1981), and enter data (Collins 1970; Rustemeyer 1977; Fowler and Mangione 1990, p.82). These interviewer-specific behaviors introduce correlations among responses within each interviewer's workload. Since responses *within* each workload tend to be similar, expected mean responses *between* interviewers tend to be different, assuming a design where samples are randomly assigned to interviewers, i.e., an interpenetrated sample (Mahalanobis 1946), and a 100% response rate. The intra-interviewer correlation coefficient is given by:

$$\rho_{int} = \frac{between\text{-}interviewer\ variance}{between\text{-}interviewer\ variance + within\text{-}interviewer\ variance} \tag{3.1}$$

While values of ρ_{int} are small - typically less than 0.02 (Groves 1989, p.318) - the variance of a mean is inflated by $1 + (MeanWorkload - 1)\rho_{int}$, potentially leading to a substantial drop in precision when mean workloads are large (Elliott and West 2015). The term 'interviewer effects' in the literature typically refers to this increase in variance on account of interviewer measurement error (Davis et al. 2010), though, depending on context, the term can also refer to the biasing effect of interviewer behavior.

One method that has been advocated to reduce interviewer effects is to adopt 'standardized interviewing' (Fowler and Mangione 1990) or 'programmed interviewing' (Blair 1980, Cannell et al. 1975) that seeks to train all interviewers to perform their tasks in the same way so that it becomes inconsequential as to which specific interviewer conducts the interview. But along with these training procedures, survey organizations also need systems to monitor interviewer behavior "to ensure that interviewers follow procedures for standardized interviewing and do not deviate from the interview script" (Currivan et al. 2006; Tarnai 2007) and to provide quick feedback on the quality of interviewing (Chapman and Weinstein 1988; Couper et al. 1992).

3.1.1 Summary of the monitoring process

The literature places an emphasis on 'direct' monitoring that collects information on the actual interaction of the interviewer with the respondent (Fowler and Mangione 1990). For Computer Assisted Telephone Interviewing (CATI) surveys, direct monitoring can be done by listening to interviews or recording the interviews and coding them later. With technological advances such as Computer Audio-Recorded Interviewing (CARI) (Biemer et al. 2000; Thissen et al. 2008; Mitchell et al. 2008; Thissen 2014), recording has become less obtrusive, survey managers have more control such as allowing specific slices of the interview to be recorded, and storing and retrieving interviews has become easier. Coding recorded interviews is typically undertaken at an item level by trained coders who typically follow standard coding schemes mentioned in the literature (e.g., Hyman 1954, Cannell et al. 1968, Marquis 1969, Cannell et al. 1975, Mathiowetz and Cannell 1980, and Ongena and Dijkstra 2006). In their study of the behavior coding literature, Ongena and Dijkstra (2006) find 134 categories of interviewer behavior that have been used; most common coding schemes track, at a minimum, question asking, probing, interactions with respondents, and recording responses.

3.1.2 Evaluating the utility of interviewer monitoring

Behavior coding data give the survey manager a quantification of interviewing quality that can be used to measure the extent to which interviewers adhere to protocol. However, little evidence exists of whether these systems are serving their fundamental purpose: the reduction of survey error.

Two studies have approached this issue. Groves and Magilavy (1986) monitored interviewer behavior in a survey with an interpenetrated sample where 1918 respondents were interviewed by 33 interviewers. Interviewer effects for 25 variables were computed and an-

alyzed for associations with two question asking behaviors. Though interviewers differed widely in the behaviors, these were not found to be correlated with $\hat{\rho}_{int}$. Scatterplots of deviations of interviewer-specific means from the overall mean versus interviewer scores from the monitoring data did not show any pattern. Interviewer effects were also not correlated with response rate, productivity, and supervisor evaluations.

Fowler and Mangione (1990) computed $\hat{\rho}_{int}$ for 65 items where 100 interviews were conducted by 57 interviewers and eight interviewer behaviors were coded. In line with the findings in Groves and Magilavy (1986), incorrect asking behavior was not correlated with $\hat{\rho}_{int}$. Four behaviors were positively correlated with $\hat{\rho}_{int}$ (and were statistically significant at a 0.05 level), with correlation coefficients ranging from 0.2 to 0.49. Of these four behaviors, three were probing behaviors (the highest correlation was for 'failure to probe') and one was to do with incorrect recording of responses.

The Groves and Magilavy (1986) study only looks at asking behavior while the Fowler and Mangione (1990) study has a limited sample size. There have been significant improvements in computing since the 1990s that allow researchers to fit models that approximate interpenetrated samples (e.g., Hox 1994; West et al. 2013; Beullens and Loosveldt 2016), allowing for more in-depth item-level analyses. But there is little research that links behavior coding data to interviewer effects, especially in the context of production surveys (compared to laboratory-like studies). This shortcoming has been alluded to in the literature:

- "Unfortunately, although interaction coding can measure compliance with training standards, there have been few empirical links made between violating those standards and measurement error."(Groves 1989, p.387)
- "we outline a variety of procedures and techniques designed to maximize the consistency of interviewing but the test of the efficacy of these solutions lies in assessments of whether or not interviewers are affecting answers." (Fowler and Mangione 1990, p.25)

This is the gap we address in this chapter, i.e., evaluating the utility of interviewer monitoring in the pursuit of reducing interviewer measurement error.

3.2 Research questions

We define 4 specific research questions in this chapter. Our first question deals with the between-interviewer variance term of equation 3.1:

1. Do interviewing quality indicators explain estimates of interviewer response vari-

ance?

Our next question does not concern itself with the between-interviewer variance component as much as with case-level associations of quality indicators with substantive responses. Take a situation where interviewers in a survey behave neutrally for most part except for a small proportion of cases for each interviewer where there is a biasing effect, e.g., encounters with hostile respondents where interviewers might be wary of probing. Monitoring evaluations of interviewing quality will flag these situations as 'failure to probe'. But if the biasing effect for a large majority of interviewers is in the same direction, the between-interviewer component will tend to remain the same thereby masking the biasing effect. This leads us to ask:

2. Is there an association between an item's substantive response and the quality of interviewing that elicited that response?

The first two questions were about the substantive responses. Our last two questions turn to the issue of item non-response. In a study on interviewers' approach to administering socially uneasy questions, Sudman et al. (1977) do not find interviewer effects for the substantive responses. However, they find higher non-response rates for interviewers who felt the questions were sensitive to respondents. In a monitoring process, such behaviors will be captured when coders find interviewers, say, rushing through the question or not probing when required. The literature shows that interviewers vary in their ability to garner responses reflecting in interviewer-varying item missing rates (Bailar et al. 1977; Loosveldt and Beullens 2014). This leads us to ask the next two questions which mirror the questions concerning substantive responses:

3. Are interviewing quality indicators for an item predictive of estimates of non-response interviewer variance for that item?

4. Are item-level interviewing quality indicators predictive of item non-response?

3.3 Study survey

We used data from the 2015 wave of the Panel Survey of Income Dynamics (PSID) for our research. The PSID is a nationally representative survey of families and individuals in the U.S., conducted via Computer Assisted Telephone Interviewing (CATI). The survey consists of biennial waves where one respondent per family is administered a 'main interview'; supplemental studies are added to this main interview e.g., the 'transition into adulthood supplement' is asked to individuals when they become 18 years of age. Between March-December 2015, 9048 respondents were interviewed by 96 interviewers with

a response rate of 89% (calculated with respect to the previous wave). Interviews are largely conducted by telephone; only 2.8% interviews had to be conducted in face-to-face mode. An interview lasted for 80 minutes on average. The detailed questionnaire [1] and codebook [2] are available on the PSID website.

Before the survey commences, PSID interviewers undergo video training [3] on the study terminology, concepts, and individual sections. This is followed by an in-depth in-person training at Ann Arbor, Michigan, USA. Approximately 60% of interviewers (61 of 96 interviewers) in PSID 2015 were also interviewers for at least one of the previous two waves.

For this paper, we are interested in PSID's substantive data and interviewing evaluation data. These are explained below.

3.4 Data

3.4.1 Substantive data

The PSID main interview begins by taking consent from the respondent followed by questions about the family composition and member details. These 'coverscreen' questions are followed by substantive questions. On average, a respondent answers about 360 substantive questions across 11 sections as shown in Table 3.1; sections concerning employment (sections BC/DE), expenditures (section F), and health (section H) account for close to 60% of interview duration.

In 2015, to aid timely research into the aftermath of the 2008 recession, PSID released data on 357 items concerning mortgage distress, housing, food security, wealth, and computer use (belonging to Sections A, F, and W) within a month of fieldwork completion. These early release data (ER data) do not include data from 'split-off' families (split-off families consist of either a person or group of people who moved out from an existing PSID family since the prior wave's interview to form a new, economically independent family unit living in a separate housing unit). This reduces the total ER data sample size to 8262 families. However, since these data were quickly released, they did not undergo the usual editing, cleaning, and imputing processes which is an advantage for our analytical goals since data 'as collected' would better reflect interviewer effects. We therefore used ER data for these families and the regular public use file data (that have undergone all the

[1] ftp://ftp.isr.umich.edu/pub/src/psid/questionnaires/q2015.pdf
[2] ftp://ftp.isr.umich.edu/pub/src/psid/codebook/fam2015er_codebook.pdf
[3] https://psidonline.isr.umich.edu/videos.aspx

No.	Section	Substantive area	Average # items administered in an IW	Average IW duration (mins)
1	A	Housing, Utilities, Computer Use	36	7
2	BC, DE *(including EHC[1])*	Employment	46	22
3	F	Expenditures	52	11
4	G	Current income; Other family unit member education	48	9
5	R	Off-year income and public assistance	12	2
6	W	Wealth and active savings	22	5
7	P	Pensions	13	3
8	H	Health	96	14
9	J	Marriages and Children	11	1
10	KL	New head and spouse/partner background	14	3
11	M	Philanthropy	9	2
		Average interiew	**358**	**80**

1. EHC: Event History Calendar

Table 3.1: PSID substantive section descriptions.

data processing) for the remaining respondents. We used only respondents interviewed via CATI for our analysis.

3.4.2 Interviewing evaluation data

In 2015, PSID recorded two of the first four interviews in every interviewer's workload followed by a further 10% random sample, resulting in 1120 recorded interviews. A 'capture list' dictated which item in the interview was to be recorded; these items were chosen on the basis of substantive importance. For the first three weeks of fieldwork, the capture list inadvertently contained 1157 items belonging to a pretest version. This was corrected and the list pared down to 382 items. We only consider items from the 382 items for our analyses.

Of the 1120 interviews, 594 CATI interviews (53% of all the recorded interviews) were listened to by nine quality control (QC) evaluators. Owing to issues such as bad recordings or missing interviewer characteristics, only 555 interviews were available for analysis. These interviews were conducted by 92 interviewers (96% of the 96 interviewers), with a median of 6 evaluated interviews per interviewer (first quartile: 4 interviews, third quartile: 8 interviews). The recorded items accounted for a median 35% of the total

number of administered substantive items (IQR: 31% - 40%) and a median 45% of the substantive interview duration (range: 40% - 50%) within the 555 evaluated interviews.

Apart from their training and extensive experience in behavior coding, many of the QC evaluators have been interviewers themselves which especially equips them to understand interviewer behavior. An evaluator raised a QC flag for an item if she encountered an issue in any of the five interviewing dimensions in Table 3.2. QC flags were classified as 'major' or 'minor' depending on the potential impact on the substantive response.

Table 3.2: The five interviewing evaluation dimensions with sixteen categories.

No.	Interviewing dimension	Categories
1	Question asking	Altered wording; Skipped question; Question delivery; Not verbatim; Other reading error
2	Probing and clarifying	Failure to probe or clarify; Inappropriate, evaluative, or directive probe; Other probing error
3	Data entry	Wrong category; Wrong entry
4	Feedback	Emotive feedback; Other feedback error
5	Other reasons	Unprofessional conduct; Consent error; Household composition; Other error

Coders also explicitly noted the cause of a major flag; four of the sixteen categories accounted for 70% of all major flags: failure to probe or clarify (44%), altered wording (11%), inappropriate, evaluative, or directive probe (9%), and 'other entry error' (6%). The large proportion of flags due to improper probing is expected since interviewers find it the hardest skill to learn (Fowler and Mangione 1990, p.44); Hicks et al. (2010) find that interviewers probed only in 57 percent of the instances when a probe was needed.

Based on these data, we created the following three 'flag variables' to use in our models:

1. $QCFlag$, a case-level variable that indicates one of the three interviewing evaluation outcomes: 'Major flag', 'Minor flag', and 'No flag'. Flag counts are generally small at the item level (see Table 3.3 below) preventing us from splitting 'Major flag' by specific interviewer behaviors.

2. $ItemFlagProportion$, an interviewer-level variable that represents the proportion of evaluated cases for a specific item (across interviews) whch have either a major or minor flag. Since there were only a median 6 evaluated cases per item per interviewer, splitting this variable on the basis of major and minor flags (let alone specific behaviors) was not possible. Figure 3.1 plots the item flag proportions for 22 items in the PSID questionnaire (these items were chosen based on the process described below in Section 3.5.1) where we can see a fair amount of variation in these proportions across items.

3. $OverallFlagProportion$, which is the overall proportion of evaluated cases for an interviewer (across items and interviews) that have either a major or minor flag.

The median for this variable is 0.033 with values ranging from 0.008 to 0.099 (IQR: 0.021 - 0.053).

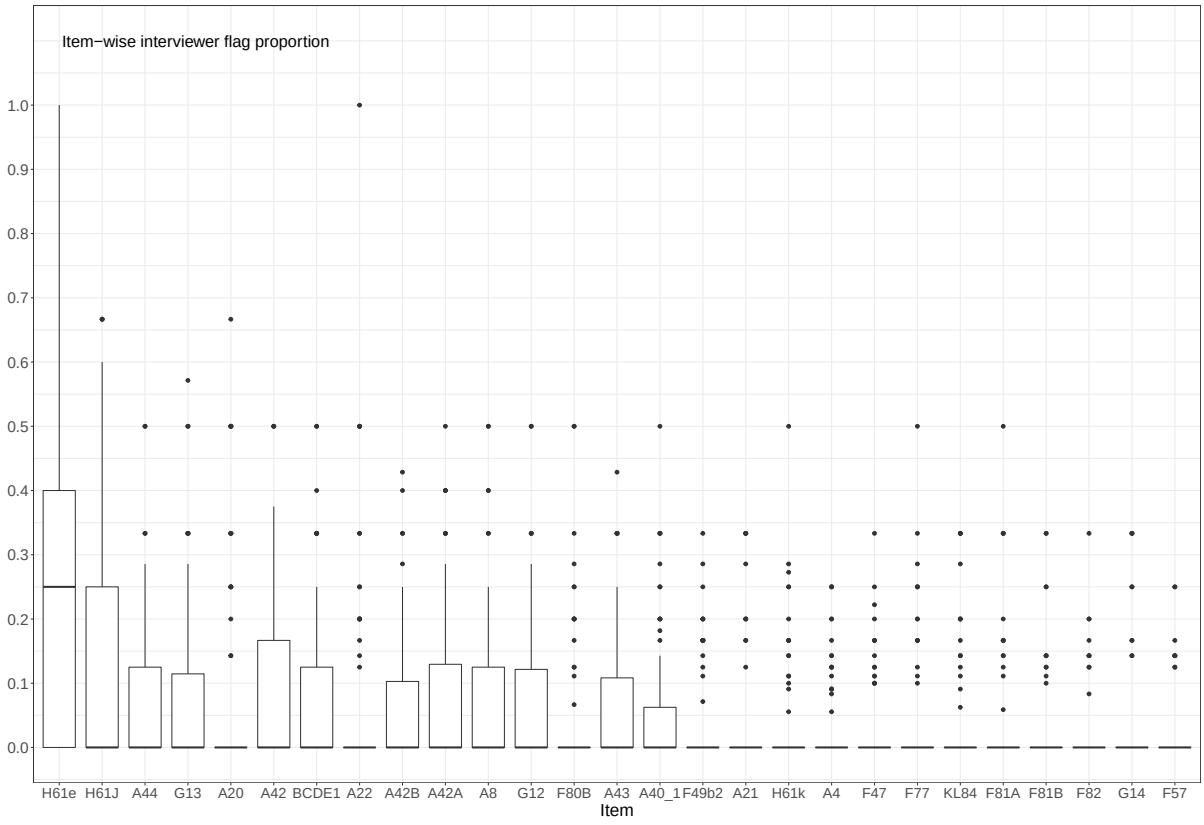

Figure 3.1: Item-wise interviewer flag proportions. Each data point in this plot is the item flag proportion for a specific interviewer. Items on the horizontal axis are sorted in descending order of the items' flag proportions seen in Table 3.3. Items with higher flag proportions have relatively more interviewers contributing to the flag proportion; items with lower flag proportions are driven by 'outliers'.

3.4.3 Interviewer characteristics

We used 3 interviewer demographic variables and 3 variables derived from interviewers' work characteristics.

1. Interviewer sex (88% of interviewers are female).
2. Interviewer age (mean: 53.6 years, standard deviation: 12.1 years)
3. Interviewer education (less than High school, 12% of interviewers; High school/GED, 35% of interviewers; some college, 28% of interviewers, college graduate and above, 25% of interviewers).
4. Interviewer workload, i.e., number of conducted interviews (mean: 114.5 interviews, standard deviation: 41.6 interviews).

5. Mean interviews per day (mean: 1.2 daily interviews, standard deviation: 0.12 daily interviews); even a moderate workload may lead to interviewer fatigue if completed in a short time period. We include partial interviews in this calculation.

6. The coefficient of variation (CV) of the number of daily interviews conducted (mean: 0.53, standard deviation: 0.11); interviewers who are more consistent with the number of daily interviews may be associated with better interviewing quality.

We refer to these 6 interviewer characteristics as 'non-flag' variables.

3.5 Methods

3.5.1 Choosing items for analysis

We started by choosing substantive items that had a minimum of 200 interviewing evaluations (roughly, a minimum of 2 evaluations each for the 92 interviewers) and at least 10 flags (so we have some potential effects to measure). We then dropped the following items: 'Event History Calendar' (EHC), which is really a battery of questions on employment and residence but which the interviewing evaluation process treats as a single 'item'; two open-ended items on the nature of work and industry (BC20 and BC21); a reading evaluation of the introduction to section M of the questionnaire; a question on which specific member is covered by health insurance (H61D3); and an item on which family member's employer provides insurance (H61f). Finally, we removed binary item 'G17f', due to a very low (only 0.2%) substantive response proportion. This yielded 27 items which are a good mix of response types: numeric (14 items, with 13 of them involving monetary values), binary (7 items), and multinomial (6 items). Seventeen items have a median of 5 or 6 evaluations per interviewer, 9 items have a median 4 evaluations per interviewer, and 3 items have a median 3 evaluations per interviewer. More item details are in Table 3.3 with items' summary statistics in Appendix 3.A.

3.5.2 Interviewer considerations

For the chosen items, we checked the number of interviewers with only one evaluation ('single-evaluation' interviewers). We found that 16 of the 27 items had up to 2 single-evaluation interviewers, 5 items had 4 or 5 such interviewers, and 3 items had 8 such interviewers. Three items (items A20, A21, and A22, that also had the least median number of evaluations) had 12-13 such interviewers. Such single-evaluation interviewers would be associated with very imprecise estimates of $ItemFlagProportion$. Also, at

least two evaluations are required to compute a within-interviewer variance. We therefore removed these single-evaluation interviewers. Table 3.3 shows a fairly large pool of interviewers even after removing these interviewers.

Table 3.3: Description of the 27 analysis items. Data are sorted by 'Flag %'. The % Major flag' column is computed on the base of '# Flags'. Item descriptions are not precise; please see the questionnaire for the detailed question. The '$' sign in the 'Response type' column signifies that a monetary amount is required as response to the question.

No.	Item	Description	Response type	Interview evaluation data					Substantive data	
				# IWs	# IWERs	# Flags	Flag %	% Major flags	# IWs	Non-response %
1	H61e	Person 1 type of health insurance	Multinomial	467	89	125	26.7%	53%	6999	0.8%
2	H61J	Monthly health insurance amount	Numeric ($)	348	81	46	13.3%	78%	4390	10.7%
3	A44	Monthly telephone and internet expenses	Numeric ($)	567	90	45	7.9%	64%	7640	1.0%
4	G13	Head's annual gross wages/salaries	Numeric ($)	406	84	31	7.5%	70%	6165	5.7%
5	A42	Utility bill type	Multinomial	567	90	41	7.2%	73%	7850	0.6%
6	BCDE1	Person 1 employment status	Multinomial	567	90	41	7.2%	70%	8060	0.3%
7	A20	Present home value	Numeric ($)	286	74	21	7.0%	48%	4027	2.9%
8	A42B	Monthly electricity expenses	Numeric ($)	468	88	32	6.8%	61%	5939	1.9%
9	A42A	Monthly gas expenses	Numeric ($)	483	88	33	6.8%	77%	4392	3.1%
10	A22	Total annual homeowner's insurance	Numeric ($)	283	73	20	6.8%	42%	3470	14.9%
11	A8	Number of rooms	Numeric	567	90	37	6.5%	62%	7865	0.4%
12	A43	Monthly water and Sewer expenses	Numeric ($)	560	90	35	6.2%	52%	5162	3.7%
13	G12	Did Head earn any wages or salary beside uninc. business?	Binary	567	90	34	6.2%	79%	8620	0.1%
14	F80B	Monthly car gasoline expenses	Numeric ($)	342	79	19	5.4%	63%	7025	0.0%
15	A40_1	Home heating type (1st mention)	Multinomial	567	90	30	5.3%	69%	8631	0.7%
16	F49b2	Vehicle type	Multinomial	419	86	18	4.3%	44%	6546	0.3%
17	A21	Total annual property tax	Numeric ($)	283	73	12	4.1%	18%	3738	8.6%
18	H61k	Did anyone at home go without insurance in the past?	Binary	567	90	22	3.9%	73%	8593	0.5%
19	F81A	Monthly bus and train fare expenses	Numeric ($)	402	83	14	3.4%	83%	8381	0.0%
20	A4	Dwelling-unit type	Multinomial	567	90	19	3.3%	89%	7898	0.0%
21	F47	Any vehicle owned/leased for personal use?	Binary	567	90	19	3.3%	68%	8625	0.1%
22	KL84	Spouse attending regular school?	Binary	503	89	16	3.2%	69%	6872	0.1%
23	F77	Monthly car insurance amount	Numeric ($)	463	88	15	3.2%	60%	6738	4.5%
24	F81B	Monthly taxicab expenses	Numeric ($)	402	83	11	2.7%	100%	8395	0.0%
25	F82	Any school-related expenses	Binary	402	83	10	2.4%	70%	8608	0.1%
26	G14	Did Head earn any additional income (e.g. tips)?	Binary	406	84	10	2.4%	20%	6512	0.6%
27	F57	Vehicle used for personal purposes?	Binary	463	88	10	2.2%	30%	7174	0.1%

3.5.3 Model: Do quality indicators explain interviewer response variance?

Research question 1 is about exploring possible associations between the QC flag variables and interviewer response error variance. Consider the following model with interviewer-varying intercepts fit to y_{ij} responses of a certain continuous item. For all models in this chapter, the subscript j refers to a respondent who is interviewed by interviewer i.

$$y_{ij} = \alpha_0 + u_{0i} + \mathbf{X_{ij}^T}\boldsymbol{\alpha_X} + \epsilon_{ij} \tag{3.2}$$

$$u_{0i} \overset{iid}{\sim} N(0, \sigma^2_{iwer})$$

$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2_e)$$

$$u_{0i} \perp \epsilon_{ij}$$

In this model (and all others in this chapter), $\mathbf{X}$ is a vector of respondent covariates via which we seek to approximate an interpenetrated design; we do not have substantive interest in them. We included three household-level variables and three individual-level variables in $\mathbf{X}$: number of adults in the family (1, 2, 3, and 4+ adults), was included to account for possible higher income and expense response values in larger families. Also, many individual-level questions in the PSID questionnaire are repeated for every adult in the home, potentially adding to respondent and interviewer burden and therefore increasing response error; the number of children at home (0, 1, 2, 3, and 4+ children) was included to account for potentially larger dwelling units and more expenses.; reported 2014 household income ($\leq$ \$25K, \$25K − \$50K, \$50K − \$75K, \$75K − \$100K, and $> 100K$) was included since the survey is primarily economic in nature and many response values would be correlated with the household economic condition; sex was included since past research suggests that this may be associated with recall accuracy (Skowronski and Thompson 1990; Auriat 1993) and measurement error on economic data (Lee and Lee 2012); education (some high school or less, high school graduate, come college, and college graduate and above) was included since this is known to be correlated with cognitive sophistication (Krosnick and Alwin 1987, Krosnick1991); and finally, age ($\leq$ 25 years, 25-34 years, 35-54 years, 55-64 years, 65-74, 75 years and above) was included since it is known to account for response effects even after taking into account respondent education (Knauper 1999).

The categorization of continuous control variables helped overcome initial problems during model estimation. It also ensured that outlier values do not overly influence our inferences. We checked if some control variable values were concentrated only among only a few interviewers but this was not the case. These variables also have low item

missing rates, the highest being that of total income at 1.4%.

Our interest is centered on $\hat{\sigma}^2_{iwer}$, the estimate of interviewer response variance, and its statistical significance was tested using the 50:50 χ^2 approach (Self and Liang 1987). If $\hat{\sigma}^2_{iwer}$ was significant, we fit 3 models: 'Non-flag model', where only the interviewer-level non-flag variables were added to equation 3.2; 'Flag model', where only the interviewer-level flag variables were added to equation 3.2; and 'Full model' that included both the non-flag and flag variables. The full model is shown in Equation 3.3 where $\mathbf{W}$ is the vector of non-flag variables; the non-flag and flag models have the same structure with only the relevant variables included.

$$\begin{aligned} y_{ij} =& \alpha'_0 + u'_{0i} + \mathbf{X^T_{ij}}\boldsymbol{\alpha'_X} + \mathbf{W^T_i}\boldsymbol{\alpha'_W} + \\ & \alpha'_1 ItemFlagProportion_i + \alpha'_2 OverallFlagProportion_i + \epsilon'_{ij} \end{aligned} \tag{3.3}$$

$$u'_{0i} \overset{iid}{\sim} N(0, \sigma'^2_{iwer})$$

$$\epsilon'_{ij} \overset{iid}{\sim} N(0, \sigma'^2_e)$$

$$u'_{0i} \perp \epsilon'_{ij}$$

The models in (3.2) and (3.3) were run on on all available PSID cases (rather than only the interviewing evaluation cases) giving us more precise $\hat{\sigma}^2_{iwer}$ and $\hat{\sigma}'^2_{iwer}$. The flag variables were computed from the QC evaluation data and were used as an estimate of the full sample flag proportions.

Given our research question, we wanted to evaluate the performance of the flag and non-flag variables in explaining $\hat{\sigma}^2_{iwer}$, which we measured by the proportion of variance explained, $\hat{p}_{ExplVar} = (\hat{\sigma}^2_{iwer} - \hat{\sigma}'^2_{iwer})/\hat{\sigma}^2_{iwer}$. Computing $\hat{p}_{ExplVar}$ for the three models gives us a measure of the incremental utility of flag variables over the non-flag variables in explaining $\hat{\sigma}^2_{iwer}$.

In fitting the non-flag model, we first fit models with individual non-flag variables. Then, starting with the non-flag variable that yielded the highest $\hat{p}_{ExplVar}$, other variables were added if they added at least an incremental 0.01 to $\hat{p}_{ExplVar}$.

For the flag model, we first separately fit 'Item flag' and 'Overall flag' models that contained only those particular flag proportion variables. We tried square, square root, cube, and cube root transformations of these variables to explore possible non-linear relationships with the outcome variable (equation 3.3 shows only linear terms for simplicity of representation). Higher exponent terms were accompanied by the corresponding lower terms as well, and the transformation with the maximum $\hat{p}_{ExplVar}$ was chosen (if at all better than using only the linear term). Terms from the item flag and overall flag models

were then put together to form the final flag model; Overall flag terms were included only if they typically added at least 0.01 over the $\hat{p}_{ExplVar}$ by the item flag terms.

Finally, terms from the non-flag and flag models were jointly included to form the full model in equation 3.3.

Model for a binary response item.
When y_{ij} is a binary variable with $y_{ij} \sim BER(p_{ij})$, we fit logit models where the predictor part is structurally similar to the linear models above (the same coefficients have been retained for simplicity):

$$log(\frac{p_{ij}}{1-p_{ij}}) = \alpha_0 + u_{0i} + \mathbf{X_{ij}^T}\boldsymbol{\alpha_X} \tag{3.4}$$

$$u_{0i} \overset{iid}{\sim} N(0, \sigma^2_{iwer})$$

$$\begin{aligned} log(\frac{p_{ij}}{1-p_{ij}}) =& \alpha'_0 + u'_{0i} + \mathbf{X_{ij}^T}\boldsymbol{\alpha'_X} + \mathbf{W_i^T}\boldsymbol{\alpha'_W} + \\ & \alpha'_1 ItemFlagProportion_i + \alpha'_2 OverallFlagProportion_i \end{aligned} \tag{3.5}$$

$$u'_{0i} \overset{iid}{\sim} N(0, \sigma'^2_{iwer})$$

The modeling and inferential approach is the same as in the linear model setting described above. For multinomial items, we fit separate logistic models to each category that constituted at least 5% of the responses. Doing so, rather than fitting a single multinomial model, gave us the flexibility to fit models with potentially different forms for each category. While not technically correct, for descriptive ease we refer to these separate categories as 'items' in our results; the 27 entries in Table 3.3 now yield 45 'items'.

Confidence intervals for $\hat{p}_{ExplVar}$.
To ascertain if a purported reduction in $\hat{\sigma}^2_{iwer}$ due to the added variables is not just sampling variation, we computed 95% bootstrap confidence intervals around $\hat{p}_{ExplVar}$. Among the several choices available to construct bootstrap confidence intervals (Carpenter and Bithell 2000), we chose the bias-corrected and accelerated (BCa) method (Efron 1987) since it provides an opportunity to correct for non-normality, bias, and non-constant standard error in estimates (Efron and Hastie 2016, p.193), and often provides claimed coverage probabilities even for small samples (Efron and Hastie 2016, p.192). One thousand datasets were generated from the original data by first resampling interviewers and then selecting all cases from the resampled interviewers. We then fit the final non-flag, flag, and full models to each of the 1000 datasets (for all selected items) and computed

$\hat{p}_{ExplVar}$ for these. These resample estimates were used to compute 95% BCa intervals using the steps outlined in Efron and Tibshirani 1993 (p.185-186); the R code is given in Appendix 3.B. The 'acceleration constant' required to construct the intervals were separately computed via a leave-one-interviewer-out Jackknife procedure using the formula in Leeden et al. 2008 (p. 419). If a model (e.g., the flag model) for an item encountered convergence/estimation issues in a particular resample, then estimates for all models (i.e, the non-flag and full models in this example) from that resample were discarded, ensuring a common resample base for all models. While the bootstrap approach also allows one to also compute bias-corrected estimates themselves, this can be 'often problematic' and 'dangerous in practice' given the high variability of bias (Efron and Tibshirani 1993, p. 138); we use the original sample-based $\hat{p}_{ExplVar}$ for our inferential needs.

3.5.4 Model: Associations between substantive responses and interviewing quality indicators

Research question 2 is about possible associations between case-level QC flags and substantive responses, with the null hypothesis being that there are no such associations. Then, assuming a successful approximation to interpenetration, a regression of the substantive response on interviewing quality indicators should show no statistically significant effects. However, the nature of some questions may be such that interviewing errors push measurement errors in the same direction, e.g., in the case of a complex question that has multiple conditions, a quick questioning style can systematically mislead respondents to provide an erroneous answer based only on the initial condition. Since these errors are pushed in the same direction, a between-interviewer analysis may fail to detect this effect if it occurs for many interviewers but a regression of substantive response on the quality flags will show an association.

Our model for a numeric variable is shown in equation 3.6. The random interviewer intercepts (v_{0i}) are included only to account for the nesting of respondents within interviewers - unlike research question 1, we are not interested in this component here. Statistically significant fixed effects ($\hat{\beta}_1$ or $\hat{\beta}_2$) indicate associations between the response and the quality flag/s. We also allow the coefficients to vary by interviewer to investigate if these associations are driven by specific interviewers. This is also useful in cases where the fixed effect, i.e., the average effect across interviewers, may not be significant but certain interviewer-specific effects might be significant. Covariances between the random coefficients and random intercepts are introduced to see if interviewer-level associations between the QC flags and the response are related to the mean response value obtained

by interviewers.

$$y_{ij} = \beta_0 + v_{0i} + \mathbf{X_{ij}^T}\boldsymbol{\beta_X} + (\beta_1 + v_{1i})I(QCFlag_{ij} = \text{“}Major\text{”}) + (\beta_2 + v_{2i})I(QCFlag_{ij} = \text{“}Minor\text{”}) + \zeta_{ij} \tag{3.6}$$

$$\begin{pmatrix} v_{0i} \\ v_{1i} \\ v_{2i} \end{pmatrix} \overset{iid}{\sim} N\left[\mathbf{0}, \begin{pmatrix} \tau^2_{0_{iwer}} & \tau_{01_{iwer}} & \tau_{02_{iwer}} \\ \tau_{01_{iwer}} & \tau^2_{1_{iwer}} & 0 \\ \tau_{02_{iwer}} & 0 & \tau^2_{2_{iwer}} \end{pmatrix}\right]$$

$$\zeta_{ij} \overset{iid}{\sim} N(0, \tau_e^2)$$

$$(v_{oi}, v_{1i}, v_{2i}) \perp \zeta_{ij}$$

Random effect variances/covariances were tested using the 50:50 χ^2 test (Self and Liang 1987) but relevant variance terms were retained even if these were insignificant since they make substantive sense; covariance terms, however, were not retained in such cases. In case variance/s of the random coefficients are significant, we also add interactions of the interviewer-level non-flag variables, $\mathbf{W}$, to see if they explain the variances.

$$\begin{aligned} y_{ij} =& \beta'_0 + v'_{0i} + \mathbf{X_{ij}^T}\boldsymbol{\beta'_X} + \mathbf{W_i^T}\boldsymbol{\beta'_W} + \\ & (\beta'_1 + \mathbf{W_i^T}\boldsymbol{\beta'^{(Major)}_W} + v'_{1i})I(QCFlag_{ij} = \text{“}Major\text{”}) + \\ & (\beta'_2 + \mathbf{W_i^T}\boldsymbol{\beta'^{(Minor)}_W} + v'_{2i})I(QCFlag_{ij} = \text{“}Minor\text{”}) + \zeta'_{ij} \end{aligned} \tag{3.7}$$

$$\begin{pmatrix} v'_{0i} \\ v'_{1i} \\ v'_{2i} \end{pmatrix} \overset{iid}{\sim} N\left[\mathbf{0}, \begin{pmatrix} \tau'^2_{0_{iwer}} & \tau'_{01_{iwer}} & \tau'_{02_{iwer}} \\ \tau'_{01_{iwer}} & \tau'^2_{1_{iwer}} & 0 \\ \tau'_{02_{iwer}} & 0 & \tau'^2_{2_{iwer}} \end{pmatrix}\right]$$

$$\zeta'_{ij} \overset{iid}{\sim} N(0, \tau'^2_e)$$

$$(v'_{oi}, v'_{1i}, v'_{2i}) \perp \zeta'_{ij}$$

The predictor part of models for binary response items, where $y_{ij} \sim BER(p_{ij})$, is structurally similar to the above equations. As in research question 1, we model individual categories of a multinomial items separately. The basic model for binary items is as follows (the model with the non-flag variables follows equation 3.7 similarly).

$$log(\frac{p_{ij}}{1-p_{ij}}) = \beta_0 + v_{0i} + \mathbf{X_{ij}^T}\boldsymbol{\beta_X} + (\beta_1 + v_{1i})I(QCFlag_{ij} = \text{“}Major\text{”}) + (\beta_2 + v_{2i})I(QCFlag_{ij} = \text{“}Minor\text{”}) \tag{3.8}$$

$$\begin{pmatrix} v_{oi} \\ v_{1i} \\ v_{2i} \end{pmatrix} \overset{iid}{\sim} N \left[\mathbf{0}, \begin{pmatrix} \tau^2_{0_{iwer}} & \tau_{01_{iwer}} & \tau_{02_{iwer}} \\ \tau_{01_{iwer}} & \tau^2_{1_{iwer}} & 0 \\ \tau_{02_{iwer}} & 0 & \tau^2_{2_{iwer}} \end{pmatrix} \right]$$

Since this particular analysis is an interview-level analysis, we are restricted to the interviewing evaluation data for model fitting. For this reason, the following 10 binary response 'items' ended up with very few counts in the smaller category (less than 20 cases) and were therefore not analyzed: KL84, G17f, category codes 12 and 97 of H61e, category codes 4, 10 and 97 of A40_1, category codes 1 and 4 of F49b2, and category code 6 of item A4.

3.5.5 Model: Do quality indicators explain interviewer non-response variance?

Research question 3 investigates possible associations between interviewer-level flag variables and interviewer non-response variance. Unlike research question 1 where we had to fit models for each item separately, here we make use of the fact that the form of the outcome variable is common across items, i.e., 1 = non-response, 0 = response; we fit a single model for all items using 'item' as a random effect thus enabling sharing of information across items. The non-response indicator, $y^{(NR)}_{ijk} \sim BER(q_{ijk})$, where item k occurs across interviews thus giving a crossed data structure; the superscript "NR" over y_{ijk} is to highlight that the variable denotes non-response. Our base model is as follows where we include covariances between the interviewer and item random effects since different sets of interviewers could be challenged in obtaining a response for different sets of items.

$$log(\frac{q_{ijk}}{1-q_{ijk}}) = \gamma_0 + \omega_{0i} + \omega_{0j} + \omega_{0k} + \mathbf{X^T_{ij}}\boldsymbol{\gamma_X} \tag{3.9}$$

$$\begin{pmatrix} \omega_{0i} \\ \omega_{0j} \\ \omega_{0k} \end{pmatrix} \sim N \left[\mathbf{0}, \begin{pmatrix} \eta^2_{Iwer} & 0 & \eta_{IwerItem} \\ 0 & \eta^2_{Resp} & 0 \\ \eta_{IwerItem} & 0 & \eta^2_{Item} \end{pmatrix} \right]$$

Since we are working with a single model, we could have potentially included more items than restrict ourselves to the selected 27 items. But we retained only the latter so that all models in this research are working with the same item base. Analogous to the model in equation 3.3 for research question 1, we now introduce the non-flag and flag variables.

$$log(\frac{q_{ijk}}{1-q_{ijk}}) = \gamma'_0 + \omega'_{0i} + \omega'_{0j} + \omega'_{0k} + \mathbf{X_{ij}^T}\boldsymbol{\gamma'_X} + \mathbf{W_i}^T\boldsymbol{\gamma'_W} + \gamma_1 ItemFlagProportion_{ik} + \gamma_2 OverallFlagProportion_i \tag{3.10}$$

$$\begin{pmatrix} \omega'_{0i} \\ \omega'_{0j} \\ \omega'_{0k} \end{pmatrix} \sim N \left[\mathbf{0}, \begin{pmatrix} \eta'^2_{Iwer} & 0 & \eta'_{IwerItem} \\ 0 & \eta'^2_{Resp} & 0 \\ \eta'_{IwerItem} & 0 & \eta'^2_{Item} \end{pmatrix} \right]$$

The modeling goal and steps closely follow research question 1 in terms of testing the variance components and covariances, trying out different transformations of the flag variables, and running the non-flag, flag, and full models. However, unlike research question 1 no bootstrapping was undertaken to estimate uncertainty of $\hat{p}_{ExplVar}$; this is computationally prohibitive given that there were 225,572 cases overall (cases for all the 27 items pooled together).

Item non-response in PSID is coded in two ways (Andreski et al. 2007): Don't know ('DK') and Not-ascertained/Refused ('NA/Refused'). Twenty-five of the 27 items in our analysis have separate codes for DK and NA/Refused. Although item non-response rates for PSID are small (typically less than 2%), due to our pooling approach we had a sufficient number of cases to conduct separate analyses for these two non-response categories; there were 2342 DK cases and 834 NA/Refused cases among the 225,572 cases. Shoemaker et al. (2002) show that DKs are linked with the lack of respondents' cognitive effort while refusals are associated with lack of cognitive effort as well as item sensitivity. Comparing $\hat{p}_{ExplVar}$ for overall non-response, non-response due to only DK, and non-response due to only NA/refusal can point us to the non-response mechanism/s getting captured by the QC process. The DK and NA/Refused outcome models are structurally the same as the above overall item non-response model in Equation 3.10 with the outcomes changed accordingly. The two items that had a common non-response code were included in both the 'DK' and 'NA/Refused' analyses.

3.5.6 Model: Associations between item non-response and interviewing quality indicators

Here we are checking for case-level associations between item non-response and quality flags. Despite using the pooling approach of research question 3, we were unable to estimate models for DK and NA/Refusal outcomes separately since we are using only the QC interviewing evaluation data; of the total 13,655 cases used for this analysis, there were 172 non-response cases, split as 122 for DK and 50 for NA/Refusal. However, unlike

research question 3 where we used flag variables summarized at the interviewer-level, using disaggregated case-level data allowed us to split 'Major flag' into two variables: an indicator whether it was on account of failure to probe or clarify (*ProbingFailure*, 226 cases) or due to other reasons (*OtherMajor*, 284 cases); *Minor flag* had 290 cases. Note that splitting *Major flag* could not be undertaken in research question 2 (which was also at a case-level) since the analyses there had to be undertaken item-wise. Our model is as follows.

$$\begin{aligned} log(\frac{q_{ijk}}{1-q_{ijk}}) =& \delta_0 + z_{0i} + z_{0j} + z_{0k} + \mathbf{X_{ij}^T}\boldsymbol{\delta_X} + \mathbf{W}_i^T\boldsymbol{\delta_W} + \\ & (\delta_1 + W_i^T\boldsymbol{\delta_W^{(ProbingFailure)}} + z_{1i})I(QCFlag_{ijk} = \text{“}ProbingFailure\text{”}) + \\ & (\delta_2 + W_i^T\boldsymbol{\delta_W^{(OtherMajor)}} + z_{2i})I(QCFlag_{ijk} = \text{“}OtherMajor\text{”}) + \\ & (\delta_3 + W_i^T\boldsymbol{\delta_W^{(Minor)}} + z_{3i})I(QCFlag_{ijk} = \text{“}Minor\text{”}) \end{aligned} \tag{3.11}$$

$$\begin{pmatrix} z_{0i} \\ z_{oj} \\ z_{ok} \\ z_{1i} \\ z_{2i} \\ z_{3i} \end{pmatrix} \sim N \left[\mathbf{0}, \begin{pmatrix} \nu^2_{0Iwer} & 0 & 0 & \nu_{01Iwer} & \nu_{02Iwer} & \nu_{03Iwer} \\ 0 & \nu^2_{0Resp} & 0 & 0 & 0 & 0 \\ 0 & 0 & \nu^2_{0Item} & 0 & 0 & 0 \\ \nu_{01Iwer} & 0 & 0 & \nu^2_{1Iwer} & 0 & 0 \\ \nu_{02Iwer} & 0 & 0 & 0 & \nu^2_{2Iwer} & 0 \\ \nu_{03Iwer} & 0 & 0 & 0 & 0 & \nu^2_{3Iwer} \end{pmatrix} \right]$$

Our inferential interests are similar to research question 2: we allow interviewer-varying coefficients and are primarily interested in the fixed effects of the flag variables $(\delta_1, \delta_2, \delta_3)$ and the conditional modes of the random coefficients. Our secondary goals are to see how the variances are explained by W and how the flag and non-flag variables interact in predicting non-response.

3.5.7 Assessing model fit

We undertake simulation-based diagnostics as described by Hartig (2018) for all our models. The key idea is that data simulated from the fitted model should mimic the observed data if the fitted model was correctly specified (Gelman and Hill 2006, p. 158-159). To do this, a thousand datasets are simulated from the model, conditioning on all random effects. Then, for each observation a quantile residual (Dunn and Smyth 1996) - defined as the proportion of simulated values larger than the observed value - is computed and two plots are constructed as described below.

- If there are no model fit issues, we would expect the quantile residuals across obser-

vations to be uniformly distributed. We draw a quantile-quantile plot to evaluate this; more formally, a Kolmogorov-Smirnov test is conducted to detect deviation from uniformity.

- The quantile residuals are plotted against the mean simulated value for each observation (similar to the diagnostic plot of residuals versus fitted values constructed for a linear model). The mean simulated values are rank transformed and scaled to make it easier to spot issues. To make the analysis more concrete and help protect against missing patterns visually (especially when there are a lot of observations), a quantile regression is conducted between the 25th percentile, median, and 75th percentile of the mean simulated values and the quantile residuals; the quantile regression lines should ideally match horizontal lines at these percentiles that would indicate no association between the residuals and the mean simulated responses. The quantile regression is conducted using quantile regression neural network models via the QRNN package (Cannon 2011) in R, so as to be able to spot potential non-linearities in the patterns.

R code for undertaking the simulation-based analyses was adapted from the source code of the DHARMa package (Hartig 2018) and is included in Appendix 3.D.

3.5.8 Other analysis details

All models were fit with the *lme4* package (Bates et al. 2015) in the R software (Team 2013). Linear models are fit using Restricted Maximum Likelihood (REML) via the Laplace approximation and the BOBYQA optimizer. Logistic models in *lme4* use ML estimation. Here, we used the 'NLopt' implementation of the BOBQYA optimizer (Powell 2009) via its R interface 'nloptr' (Ypma et al. 2014); some testing showed that variance estimates were almost exactly the same as when the default BOBYQA optimizer was used but with more than a 50% reduction in runtime. The reduction in runtime was especially critical for research question 1 where we were undertook the computationally intensive bootstrapping.

All tests were conducted at a 5% level of significance unless noted otherwise. All numeric input variables were centered and scaled. We do not screen for 'outliers' when running our models since these are actually potentially valuable in being responsible for interviewer effects. For items H61J and F77, survey responses were obtained in different time units - follow-up questions asked about the time unit being used to report those spends. We used these time-unit responses to convert all substantive responses to a monthly figure.

Our models do not add item characteristic covariates since our focus is on the interviewer.

We do not include survey weights since we are interested in uncovering interviewer effect structures conditional on the current PSID design, not averaged over the design to the population. All our models use the vector of respondent-level covariates $\mathbf{X}$ but only as controls; effects for these are not presented in our results since they are not our focus.

Approximately 3% of all interviews were undertaken by multiple interviewers. For these interviews, we used keystroke paradata to match an item to the interviewer who actually asked the question. In the case of some items (such as A42A, utility expenses), a code of zero could mean either an actual zero value or that the item was not administered due to the skipping pattern; matching records to the keystroke paradata files also helped distinguish between these two situations.

When lowest-level covariates are added to logistic models, the underlying latent variable distribution is rescaled due to which the estimates need to be rescaled back to ensure comparability (Snijders and Bosker 1999, p. 228-229; Hox 2010, p.134; Austin and Merlo 2017). Research questions 1 and 3 entail comparisons between the flag, non-flag and full models. However, for these questions, all covariates added over the base model are at the interviewer level and not the case-level which obviates the need for such rescaling (Hox 2010, p.138).

3.6 Results

3.6.1 Research question 1 - Do quality indicators explain interviewer response variance?

Of the 45 base models, the interviewer variance component was significant for all but 6 items (F81B, F80B, G13, F82, KL84, A4_2; 'A4_2' indicates category code 2 of multinomial item A4). Further model fitting and bootstrap analyses were undertaken for the remaining 39 items. Power transformations of the flag variables were found to be much more effective in explaining $\hat{\sigma}^2_{iwer}$ compared to only the linear terms; Table 3.5 described later shows that almost all flag models with a significant $\hat{p}_{ExplVar}$ used a cube transformation.

Thirty-five of the 39 items had at least 950 valid bootstrap estimates. Of the remaining 4 items, 3 items (A40_1_10, F49b2_5, and G14) had at least 900 valid estimates while 1 item (F57) ended up with only 702 valid estimates. Bootstrap $\hat{p}_{ExplVar}$ distributions were plotted and examined for all 39 items; no concerning features such as discontinuities were seen. Since all interviewers do not conduct the same number of interviews, resample sizes were not constant. However, the coefficients of variation were small (range: 3.3% - 4.2%

across items).

Figure 3.2 summarizes $\hat{p}_{ExplVar}$ for the 39 items for each of the three models. We find a modest median $\hat{p}_{ExplVar}$ for the non-flag and flag models with the former higher than the latter (19% and 15% respectively). In comparison, the median $\hat{p}_{ExplVar}$ for the full model is at a relatively much higher 30% suggesting that the flag and non-flag variables complement each other.

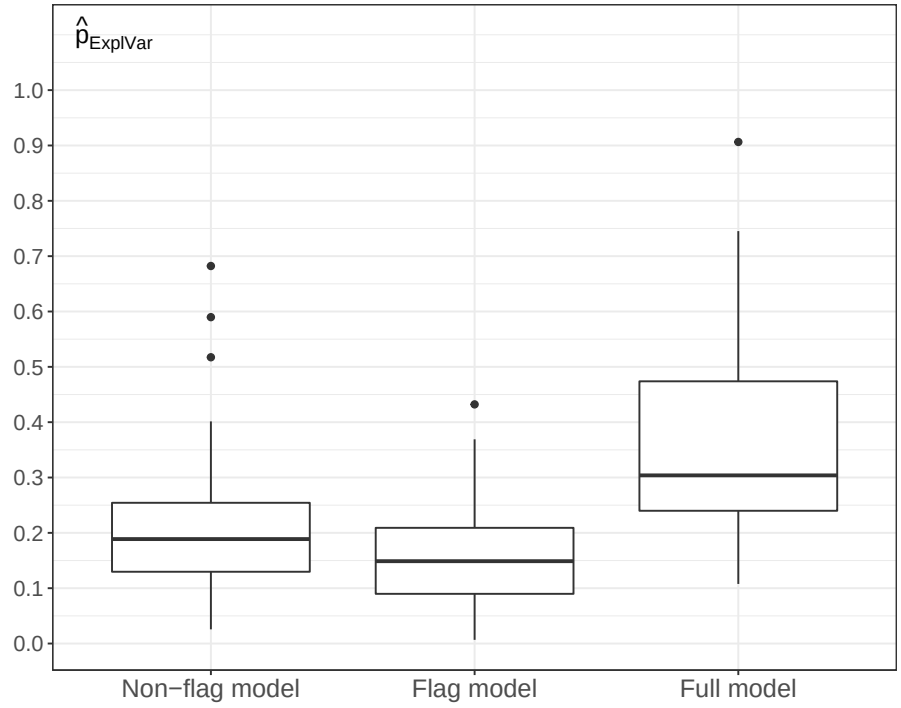

Figure 3.2: Boxplot of $\hat{p}_{ExplVar}$ for the non-flag, flag and the full models. These are computed based on the 39 analyzed items. The plot suggests the complementarity of the flag and non-flag variables in explaining $\hat{\sigma}^2_{iwer}$ in the full model.

To explore this phenomenon more and quantify the uncertainty in $\hat{p}_{ExplVar}$, Figure 3.3 plots the sorted item-wise $\hat{p}^{Flag}_{ExplVar}$ estimates (shown by circles) along with the 95% BCa confidence intervals (the superscript 'Flag' in $\hat{p}^{Flag}_{ExplVar}$ indicates that these are flag model estimates). Fourteen of the 39 items have statistically significant estimates (black circles) and the wide confidence intervals suggest that a few interviewers may be driving these effects. The plot also displays $\hat{p}^{NonFlag}_{ExplVar}$ (by 'x's; confidence intervals are not plotted for these to avoid clutter) and we find 17 significant estimates (bold 'x's). Approximately half of these estimates (9 estimates) are significant where $\hat{p}^{Flag}_{ExplVar}$ are not significant; similarly, of the 14 significant $\hat{p}^{Flag}_{ExplVar}$, 6 belong to items where $\hat{p}^{NonFlag}_{ExplVar}$ are non-significant.

Figure 3.4 now plots the sorted $\hat{p}^{NonFlag}_{ExplVar}$ along with the BCa confidence intervals onto which are overlaid $\hat{p}^{Full}_{ExplVar}$ (shown by squares). A large majority of the items (28 items) now have significant $\hat{p}^{Full}_{ExplVar}$ and we can see the incremental variance explained by the flag variables over the non-flag variables - especially for items where $\hat{p}^{NonFlag}_{ExplVar}$ are non-

significant. The flag variables explain an incremental median 17% points (IQR: 10% - 22%) over the non-flag variables for these 28 items resulting in the full model explaining a fairly sizable median 37% of $\hat{\sigma}^2_{iwer}$ (IQR: 29% - 53%). These results show that the flag and non-flag variables are indeed complementary to each other; the flag variables seem to be deriving their explanatory power by capturing interviewing behaviors that are not fully explained by interviewers' sex, age, education, and work-related characteristics.

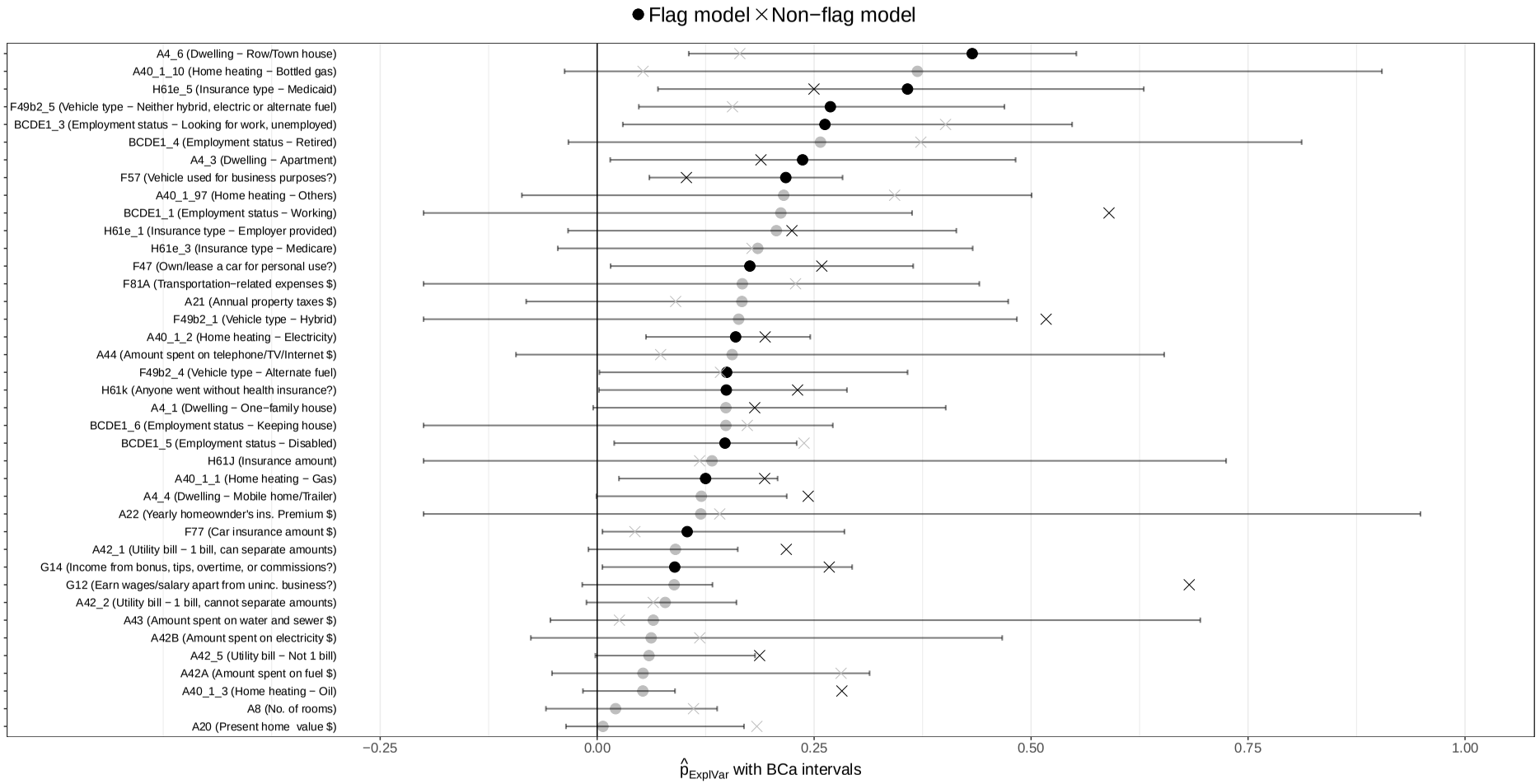

Figure 3.3: Item-wise proportions of variance explained for the flag and non-flag models. The $\hat{p}_{ExplVar}$ for the flag model are shown by circles and those from non-flag model by (x). The item descriptions are in the vertical vertical axis. In the case of multinomial items, the specific category being modeled is given as the category number following the underscore after the item name. Estimates are sorted by flag model values along with their 95% BCa confidence intervals (horizontal bars). Significant estimates are shown in bold.

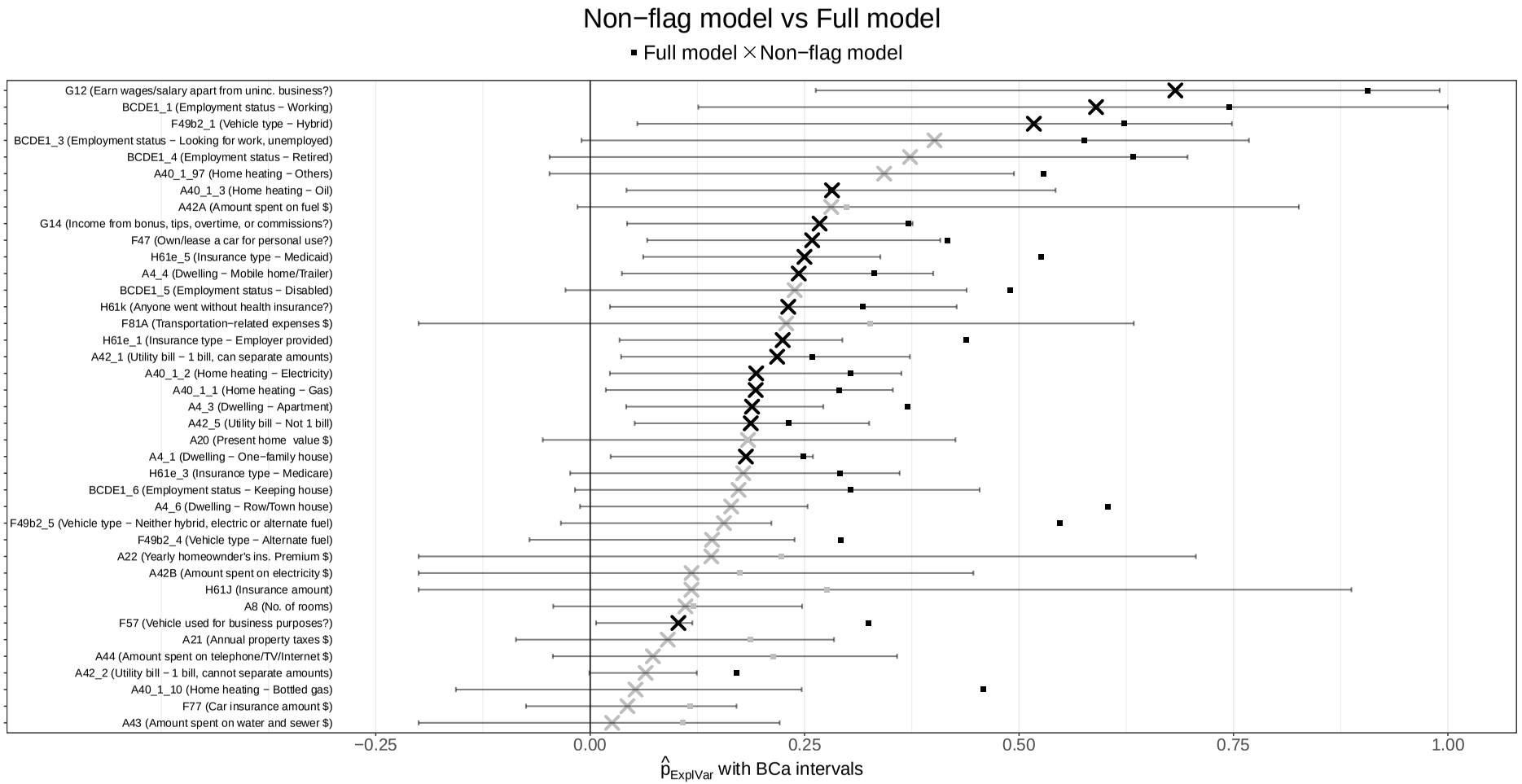

Figure 3.4: Item-wise proportions of variance explained for the non-flag and full models. The $\hat{p}_{ExplVar}$ for the non-flag model are shown by 'x' and those from the full model by squares for the 39 analyzed items (vertical axis). In the case of multinomial items, the specific category being modeled is given as the category number following the underscore after the item name. Estimates are sorted by non-flag model values along with their 95% BCa confidence intervals (horizontal bars). Significant estimates are shown in bold.

Diagnostic checks did not show any problems with model fit. Table 3.4 shows the Kolmogorov-Smirnov statistics based on the quantile residual analysis for the 28 items with significant $\hat{p}^{Full}_{ExplVar}$. Magnitudes of all K-S statistics are small; only one item, F49b2_4 (alternate fuel vehicle) is associated with a p-value less than 0.05. The diagnostic plots for this item are shown in Figure 3.5 which does not indicate any issues with model fit. Diagnostic plots for the other items are not shown since they are similar to Figure 3.5.

Table 3.4: Test of uniformity for the quantile residuals. Testing was done using the Kolmogorov-Smirnov (KS) test. Of the 28 items with significant $\hat{p}^{Full}_{ExplVar}$, only one item (F49b2_4, Alternate fuel vehicle type; data in boldface) has a p-value less than 0.05.

Item	KS test statistic	p-value
A4_1 (Dwelling - One-family house)	0.009	0.48
A4_3 (Dwelling - Apartment)	0.007	0.82
A4_4 (Dwelling - Mobile home/Trailer)	0.011	0.29
A4_6 (Dwelling - Row/Town house)	0.009	0.49
A40_1_10 (Home heating - Bottled gas)	0.011	0.3
A40_1_1 (Home heating - Gas)	0.009	0.55
A40_1_2 (Home heating - Electricity)	0.007	0.79
A40_1_3 (Home heating - Oil)	0.009	0.44
A40_1_97 (Home heating - Others)	0.011	0.21
A42_1 (Utility bill - 1 bill, can separate amounts)	0.007	0.74
A42_2 (Utility bill - 1 bill, cannot separate amounts)	0.009	0.51
A42_5 (Utility bill - Not 1 bill)	0.011	0.21
BCDE1_1 (Employment status - Working)	0.011	0.3
BCDE1_3 (Employment status - Looking for work, unemployed)	0.012	0.16
BCDE1_4 (Employment status - Retired)	0.009	0.46
BCDE1_5 (Employment status - Disabled)	0.013	0.37
BCDE1_6 (Employment status - Keeping house)	0.017	0.36
F47 (Own/lease a car for personal use?)	0.008	0.69
F49b2_1 (Vehicle type - Hybrid)	0.006	0.96
F49b2_4 (Vehicle type - Alternate fuel)	**0.018**	**0.04**
F49b2_5 (Vehicle type - Neither hybrid, electric or alternate fuel)	0.016	0.08
F57 (Vehicle used for business purposes?)	0.007	0.91
G12 (Earn wages/salary apart from uninc. business?)	0.007	0.83
G14 (Income from bonus, tips, overtime, or commissions?)	0.006	0.96
H61e_1 (Insurance type - Employer provided)	0.009	0.58
H61e_3 (Insurance type - Medicare)	0.008	0.69
H61e_5 (Insurance type - Medicaid)	0.009	0.67
H61k (Anyone went without health insurance?)	0.008	0.58

Since we had fit separate models using the *ItemFlagProportion* and *OverallFlagProportion* variables while arriving at the final flag model, we were able to assess the relative importance of these variables in explaining $\hat{\sigma}^2_{iwer}$. Three key points emerge from the results. First, *ItemFlagProportion* generally outperforms *OverallFlagProportion*, which is not surprising given that interviewers struggle with specific items. Second, *OverallFlagProportion*, however, does not perform badly - outperforming *ItemFlagProportion* for some items. This suggests that interviewer effects are not always driven by item-specific struggles but also by general interviewing behaviors. Third, for many items, the two sets of variables complement each other in explaining $\hat{\sigma}^2_{iwer}$. These results are displayed in Table 3.5.

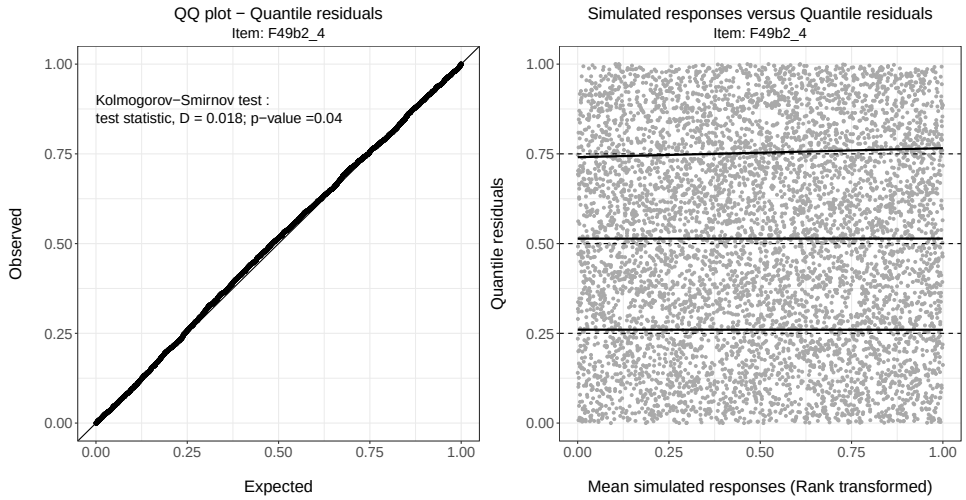

Figure 3.5: Diagnostics for the full model for item F49b2_4 (alternate fuel vehicle). The left panel compares the quantile residuals to draws from a uniform distribution. Each point in the right panel is the mean simulated response (across 1000 simulations) for an observation in the data. The solid lines in the correspond to the quantile regression lines and the dotted lines are benchmarks for these lines.

A similar analysis of the non-flag variables (Table 3.6) shows that interviewer education is the most important non-flag variable in explaining interviewer effects. This is possibly to do with the economic nature of the survey. Interviewer sex also emerges as an important variable.

While our selection of analysis items is not suited to make general inferences on item characteristics, we find that for all multinomial items except one, significant $\hat{p}_{ExplVar}$ for the first category (i.e. code = 1 in the questionnaire) were obtained only by the non-flag model (the one item had significant estimates via both the flag and non-flag models). Conversely, $\hat{p}_{ExplVar}$ were significant for the later coded categories only via the flag models. Since the later categories also tend to be the smaller ones, the flag variables are perhaps able to capture finer behavioral nuances in explaining $\hat{\sigma}^2_{iwer}$ compared to the 'coarser' non-flag variables.

Table 3.5: Comparison of item flag and overall flag proportion variables in explaining interviewer effects. Only items with significant $\hat{p}^{Flag}_{ExplVar}$ are included in the table. The last three columns of the table show the proportion of $\hat{\sigma}^2_{iwer}$ (column 3) explained by the final flag model and the subset models using only the *ItemFlagProportion* and the *OverallFlagProportion* variables. Rows are sorted in descending order of $\hat{p}^{Flag}_{ExplVar}$. No estimates of precision are provided for these estimates since the subset models were not fit to the bootstrapped datasets. For each item, the higher of the last two columns is highlighted as long as there is at least a 5% point difference between them; darker colors correspond to higher proportions. 'NA' indicates that the variable was not included in the flag model.

			Variable transformations used		$\hat{p}_{ExplVar}$		
Item	Item type	$\hat{\sigma}^2_{iwer}$	ItemFlagProportion	OverallFlagProportion	Flag model	ItemFlagProportion model	OverallFlagProportion model
A4_6 (Dwelling - Row/Town house)	Multinomial	0.37	Cube	Cube	**43%**	**33%**	**7%**
H61e_5 (Insurance type - Medicaid)	Multinomial	0.06	Cube	Cube	**36%**	**33%**	**15%**
F49b2_5 (Vehicle type - Neither hybrid, electric or alternate fuel)	Multinomial	0.22	Cube	Cube	**27%**	**17%**	**8%**
BCDE1_3 (Employment status - Looking for work, unemployed)	Multinomial	0.05	Cube	Cube	**26%**	**15%**	**12%**
A4_3 (Dwelling - Apartment)	Multinomial	0.06	Cube	Cube	**24%**	**2%**	**21%**
F57 (Vehicle used for business purposes?)	Binomial	0.10	Cube	Cube	**22%**	**4%**	**20%**
F47 (Own/lease a car for personal use?)	Binomial	0.09	Cube	Cube	**18%**	**6%**	**9%**
A40_1_2 (Home heating - Electricity)	Multinomial	0.11	Cube	Cube	**16%**	**9%**	**7%**
BCDE1_5 (Employment status - Disabled)	Multinomial	0.15	Cube	Cube	**15%**	**6%**	**10%**
F49b2_4 (Vehicle type - Alternate fuel)	Multinomial	0.56	Square	Square	**15%**	**12%**	**2%**
H61k (Anyone went without health insurance?)	Binomial	0.04	Cube	Cube	**15%**	**2%**	**9%**
A40_1_1 (Home heating - Gas)	Multinomial	0.08	Cube	Cube	**12%**	**9%**	**3%**
F77 (Car insurance amount)	Numeric	198.4	Square root	NA	**10%**	**10%**	
G14 (Income from bonus, tips, overtime, or commissions?)	Binomial	0.12	Cube	NA	**9%**	**9%**	

Table 3.6: Comparison of individual non-flag variables in explaining interviewer effects. Only items with a significant $\hat{p}_{ExplVar}^{NonFlag}$ (fourth column) are included in the table; rows are sorted by these estimates. The last six columns are the proportion of $\hat{\sigma}_{iwer}^2$ explained by individual non-flag variables. No estimates of precision are provided for these estimates since the submodels were not fit to the bootstrapped datasets. Blank cells mean that that characteristic was not included in the final non-flag model. Darker colored cells show a higher proportion of variance explained.

Item	Item type	$\hat{\sigma}_{iwer}^2$	$\hat{p}_{ExplVar}$						
			Non-flag model	Iwer sex	Iwer age	Iwer educ.	Iwer workload	Iwer mean daily contact	Iwer workload CV
G12 (Earn wages/salary apart from uninc. business?)	Binomial	0.05	68%	**33%**		**32%**	4%	3%	
BCDE1_1 (Employment status - Working)	Multinomial	0.02	59%	**23%**	2%	**25%**	5%	7%	5%
F49b2_1 (Vehicle type - Hybrid)	Multinomial	0.29	52%	4%	9%	**27%**		3%	8%
A40_1_3 (Home heating - Oil)	Multinomial	0.37	28%	**17%**		3%	2%		
G14 (Income from bonus, tips, overtime, or commissions?)	Binomial	0.12	27%		**14%**	2%	**12%**		2%
F47 (Own/lease a car for personal use?)	Binomial	0.09	26%	0.4%	6%	1%	0.1%	**13%**	7%
H61e_5 (Insurance type - Medicaid)	Multinomial	0.06	25%	8%	0.3%	2%	9%	1%	2%
A4_4 (Dwelling - Mobile home/Trailer)	Multinomial	0.33	24%	9%		**18%**			2%
H61k (Anyone went without health insurance?)	Binomial	0.04	23%	**12%**		**13%**			
A42_1 (Utility bill - 1 bill, can separate amounts)	Multinomial	0.18	22%	13%		4%			2%
H61e_1 (Insurance type - Employer provided)	Multinomial	0.18	22%	0.1%	3%	8%	0.1%	6%	3%
A4_3 (Dwelling - Apartment)	Multinomial	0.06	19%	3%		8%	2%	5%	5%
A40_1_2 (Home heating - Electricity)	Multinomial	0.11	19%	2%		**15%**			6%
A40_1_1 (Home heating - Gas)	Multinomial	0.08	19%			**17%**			4%
A42_5 (Utility bill - Not 1 bill)	Multinomial	0.13	19%	9%		6%		2%	
A4_1 (Dwelling - One-family house)	Multinomial	0.05	18%	2%		6%	5%	2%	2%
F57 (Vehicle used for business purposes?)	Binomial	0.10	10%		0.2%	7%		0%	3%

3.6.2 Research question 2 - Associations between substantive responses and interviewing quality indicators

Of the 35 items analyzed for this question, 15 items had some significant effect of interest - having either a significant flag variance component (Table 3.7) or a significant fixed effect (Table 3.8), with the exception of item A44 which is present in both tables. In addition, 3 items also had a significant random intercept variance but this is not shown in Table 3.7 since it is not the estimate of interest.

Table 3.7: Significant variances for the random 'Major flag' and 'Minor flag' coefficients. Figures in parentheses are the LR statistics and associated p-values. The '-' indicate non-significant results.

Item	**Variance Components** $\tau^2_{1_{iwer}}$ ('major flag')	$\tau^2_{2_{iwer}}$ ('minor flag')
Logistic models		
A4_2 (Utility bill - 1 bill, cannot separate amounts)	1124.5 *(7.3, 0.003)*	-
A4_3 (Dwelling - Apartment)	2.1 *(5.8, 0.008)*	-
A40_1_2 (Home heating - Electricity)	844.7 *(9.8, 0.001)*	-
G12 (Earn wages/salary apart from uninc. business?)	232.3 *(9.8, 0.001)*	-
A40_1_1 (Home heating - Gas)	675.1 *(7.2, 0.004)*	1057.1 *(4.1, 0.02)*
Linear models		
A8 (No. of rooms)	-	3.1 *(2.9, 0.04)*
H61J (Insurance type - Medicaid)	-	403919.6 *(35.7, ¡0.001)*
A44 (Amount spent on telephone/TV/Internet $)	8586.3 *(3.3, 0.034)*	

Table 3.8: Significant regression coefficient estimates for the 'Major flag' and 'Minor flag' coefficients. Figures in parentheses are 95% confidence intervals. The '-' indicate non-significant results. Logistic model estimates are presented as odds ratios.

Item	**Model coefficients** $\hat{\beta}_1$ ('major flag')	$\hat{\beta}_2$ ('minor flag')
Logistic models (odds ratios)		
A42_1 (Utility bill - 1 bill, can separate amounts)	2.6 *(1.7, 3.4)*	-
A42_2 (Utility bill - 1 bill, cannot separate amounts)	1.0 *(0.2, 1.9)*	-
A42_5 Utility bill - Not 1 bill	-3.6 *(-5, -2.1)*	-
BCDE1_3 Employment status - Looking for work, unemployed	1.7 *(0.6, 2.8)*	-
F82 (Any school-related expenses?)	3.1 *(0.9, 5.4)*	-
F49b2_5 (Vehicle type - Neither hybrid, electric or alternate fuel)	-	-2.4 *(-4.3, -0.4549)*
H61e_3 (Insurance type - Medicare)	-	-3.8 *(-6.9, -0.6456)*
Linear model		
A44 (Amount spent on telephone/TV/Internet $)	-97.2 *(-156, -38.4)*	

Proportions of variances in Table 3.7 explained by non-flag variables tended to be extreme (e.g., 100% using only interviewer education) and are therefore not shown. This is generally the result of a small number of interviewers driving the effects, i.e., the variances are more due to interviewer-specific behavioral idiosyncrasies rather than due to general interviewer demographic or work characteristic variables. We computed interviewer-specific coefficients for items which had a significant variance component in Table 3.7 by adding the conditional modes of the random effects to the fixed effects; standard errors were estimated by the square root of the sum of the variances of the fixed effect and the conditional modes (covariances between the fixed and random effects were ignored). We found significant interviewer-specific coefficients only for items A44 and H61J, plotted in Figure 3.6.

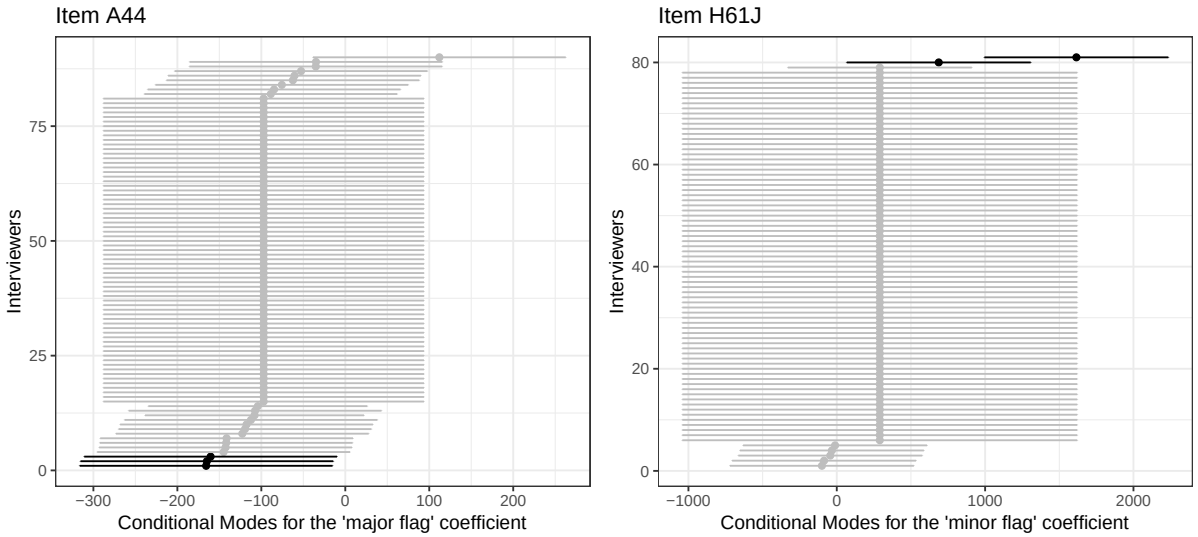

Figure 3.6: Interviewer-specific coefficients for items A44 and H61J. These are sorted coefficients with those for item A44 shown in the left panel and those from H61J shown in the right panel. The horizontal bars are the 95% confidence intervals with bold lines representing significant interviewer-specific effects.

We first focus on item A44 which asks about the monthly expense incurred on three services - telephone (including cell phone), cable and satellite TV, and Internet. The model for this item also has a significant fixed effect for the major flag (Table 3.8) that indicates that, on average, the occurrence of a major flag is associated with a $97.2 reduction in the expense reported for these services. This could indicate a primacy effect; given that there are multiple utilities involved, if the respondent simply gives the telephone expense (the first utility mentioned in the question) and this is not probed further by the interviewer, there will be an underestimate in the value obtained. Analyzing the QC data, we find that 76% of the major flags for this item are due to a failure to probe or clarify. To explore more, we drill-down to the values obtained by the three interviewers with significant interviewer-specific major flag coefficients in the left panel of Figure 3.6. We found one major flag for each for these interviewers, all of which had a recorded value of only $1, a rather small value. Of the three cases, the cause of one flag was data entry (the

interviewer might have, e.g., entered $1 instead of $100), and that of the other two were a failure to probe or clarify (the interviewer might have failed to probe the respondent even after receiving a low value for this item).

The right panel of Figure 3.6 shows two interviewers with significant interviewer-specific minor flag coefficients for item H61J, an item that asks about the monthly family insurance amount. The interviewer with the maximum conditional mode has a respondent (linked to the minor flag) whose family pays a monthly insurance premium that is 40% of the total household income (actual $ values not reported to protect confidentiality). The other interviewer has a respondent who is the sole member of the family yet pays a large sum per month on insurance. Since we are dealing with a minor flag, we do not have data on specific flag reasons but the response values seem non-ordinary.

We now turn to Table 3.8 that displays the eight models for which we obtained significant fixed effects for the flag variables. We find that the confidence intervals are wide in all cases reflecting the small flag counts that are producing the effects. Of the 8 'items' in Table 3.8, 3 belong to item A42 that asks about the type of utility bill received. A major flag increases the odds ratios of a response to category 1 (receives one bill and can separate amounts for different utilities) and category 2 (receives one bill but cannot separate amounts for different utilities) by 12.8 and 2.8 respectively compared to 0.03 for category 3 (receive different utility bills) - the odds ratio for each category are relative to the other two. Given the double-barreled nature of this question, it is likely that there could be confusion between the ability to separate utility amounts and actually getting separate bills; indeed, a check shows that 50% of the major flags for this item are due to 'wrong category'.

A major flag for item BCDE1 (a question on employment status) increases the odds ratio of category 3 ('looking for work, unemployed') being selected as compared to the other categories. Of the 8 possible response categories for this question, category 3 is the only one one where a special probing instruction is provided; if the respondent is unemployed but is *not* looking for work, category 3 is the wrong category and the interviewer has to put this into category 6 ('keeping house'). Thus, a failure to probe would result in a higher chance of category 3 being selected. The QC data align with this thinking; 52% of major flags for this item were due to 'Failure to probe or clarify' and 24% were due to 'wrong category'.

The occurrence of a major flag for item F82 (any school-related expenses incurred) increases the odds ratio of a 'yes' to the question. The actual question contains a fairly long list of possible school-related expenses; unless carefully asked, the question could be interpreted as being broad enough to answer in the affirmative. While there are only 7 major flags in all for this item in the QC data, we find that 5 major flags are to do with

improper asking or data entry issues.

The occurrence of a minor flag for item F49b2 (a question on whether the vehicle is hybrid, electric, or alternate fuel) is associated with a reduced odds ratio of category 5 ('None of the above') relative to the other categories. Examination of the questionnaire and QC data did not suggest a straightforward mechanism behind this result. Finally, the occurrence of a minor flag for item H61e (type of health insurance or coverage) reduces the odds ratio of category 3 (Medicare) being selected compared to other categories. This is the only category for the question that has a special check that the interviewer has to perform if the category is *not* selected by the respondent; a failure to probe would reduce the chances of this category occurring.

These results show that the QC indicators are at least partially successful in identifying observations that may be contributing to bias in estimates.

Figure 3.7 displays the diagnostic plots for the models whose estimates are shown in Tables 3.7 and 3.8. Looking at the QQ plot (left panel) in conjunction with the 'simulated responses versus quantile residuals' plot (right panel) for each item, in general the models show only moderate deviations from a good fit. The exceptions are items H61J (insurance amount), A44 (telephone/TV/Internet expenses), and H61e_3 (Medicare insurance); at least in case of the first two of these items, one could consider transforming the response variable but this would make interpretations more difficult.

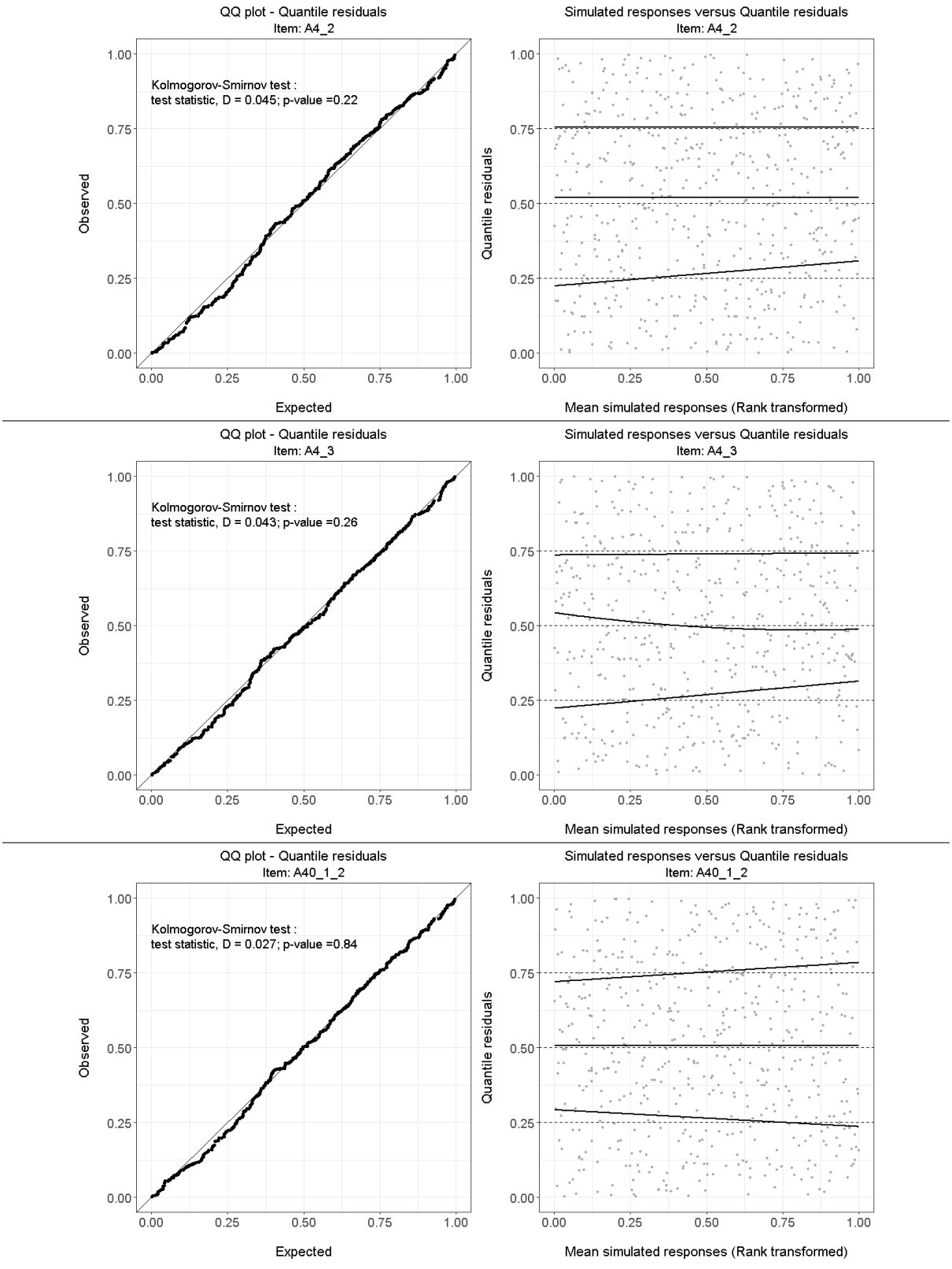

Figure 3.7: Quantile residual diagnostic plots - Models for associations between substantive values and QC indicators. Each item is on one row of the plot with the order of the items as followed in Table 3.8. The left panel compares the quantile residuals to draws from a uniform distribution. Each point in the right panel is the mean simulated response (across 1000 simulations) for an observation in the data. The solid lines in the correspond to the quantile regression lines and the dotted lines are benchmarks for these lines.

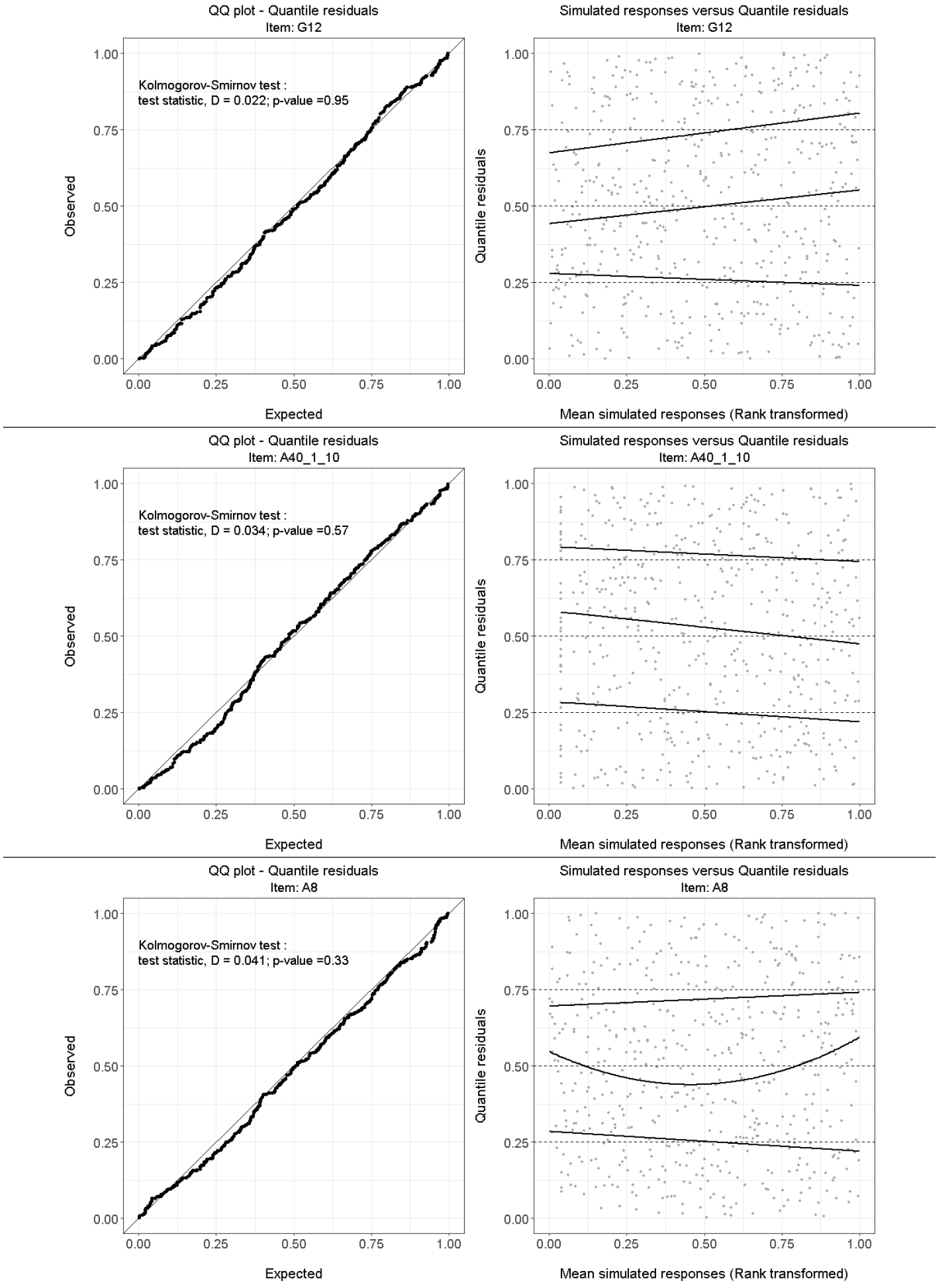

Figure 3.7 (continued).

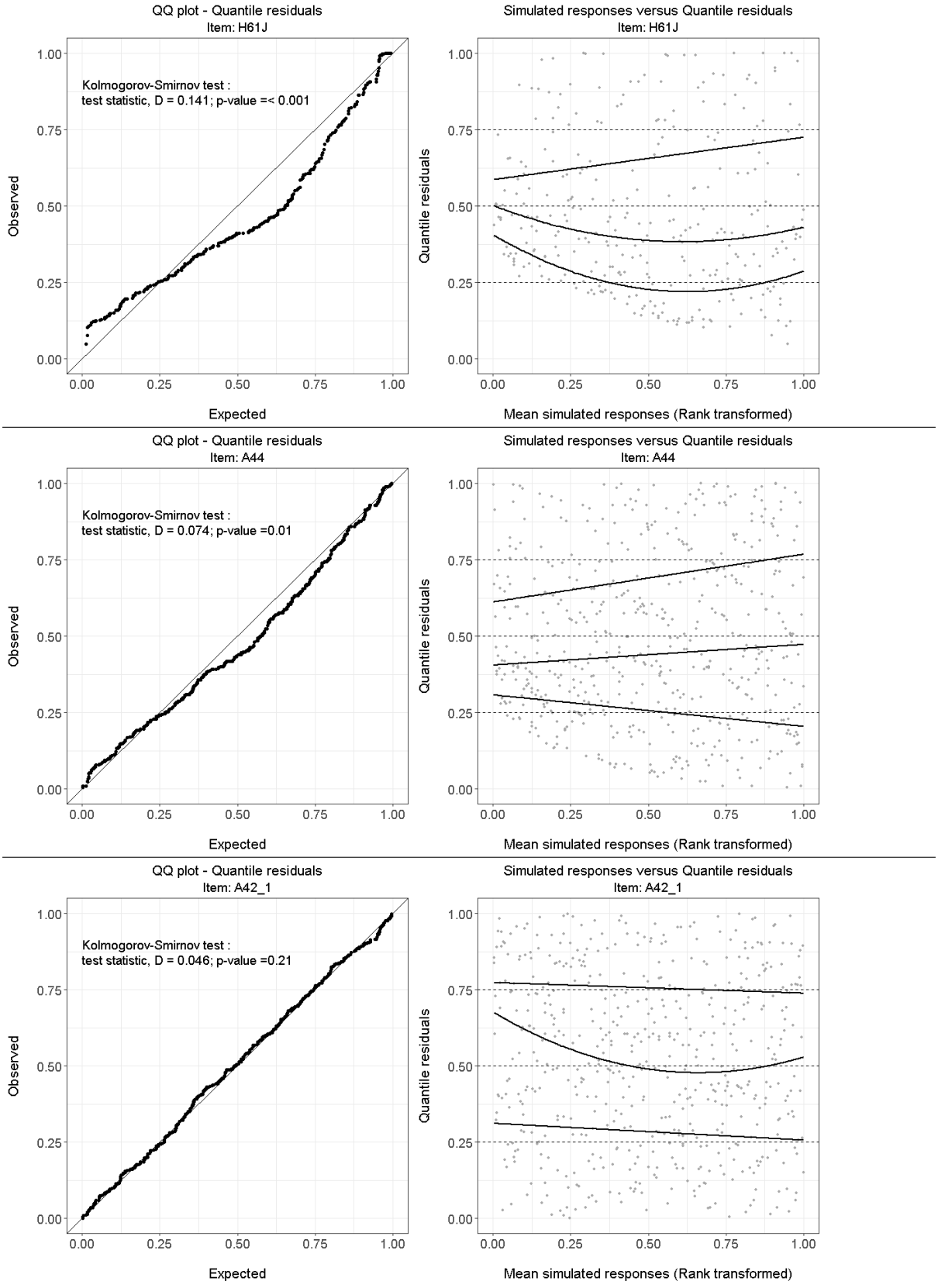

Figure 3.7 (continued).

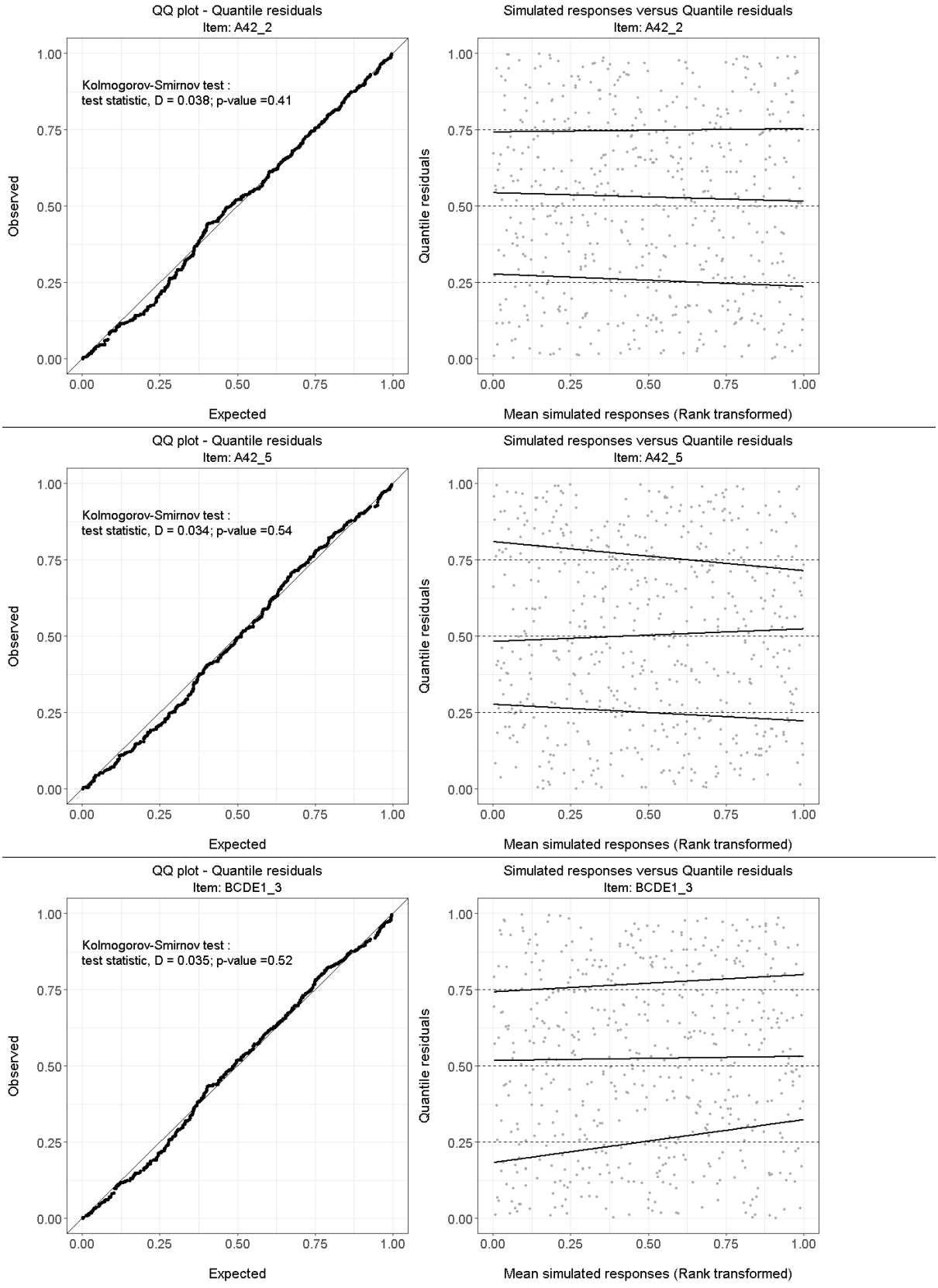

Figure 3.7 (continued).

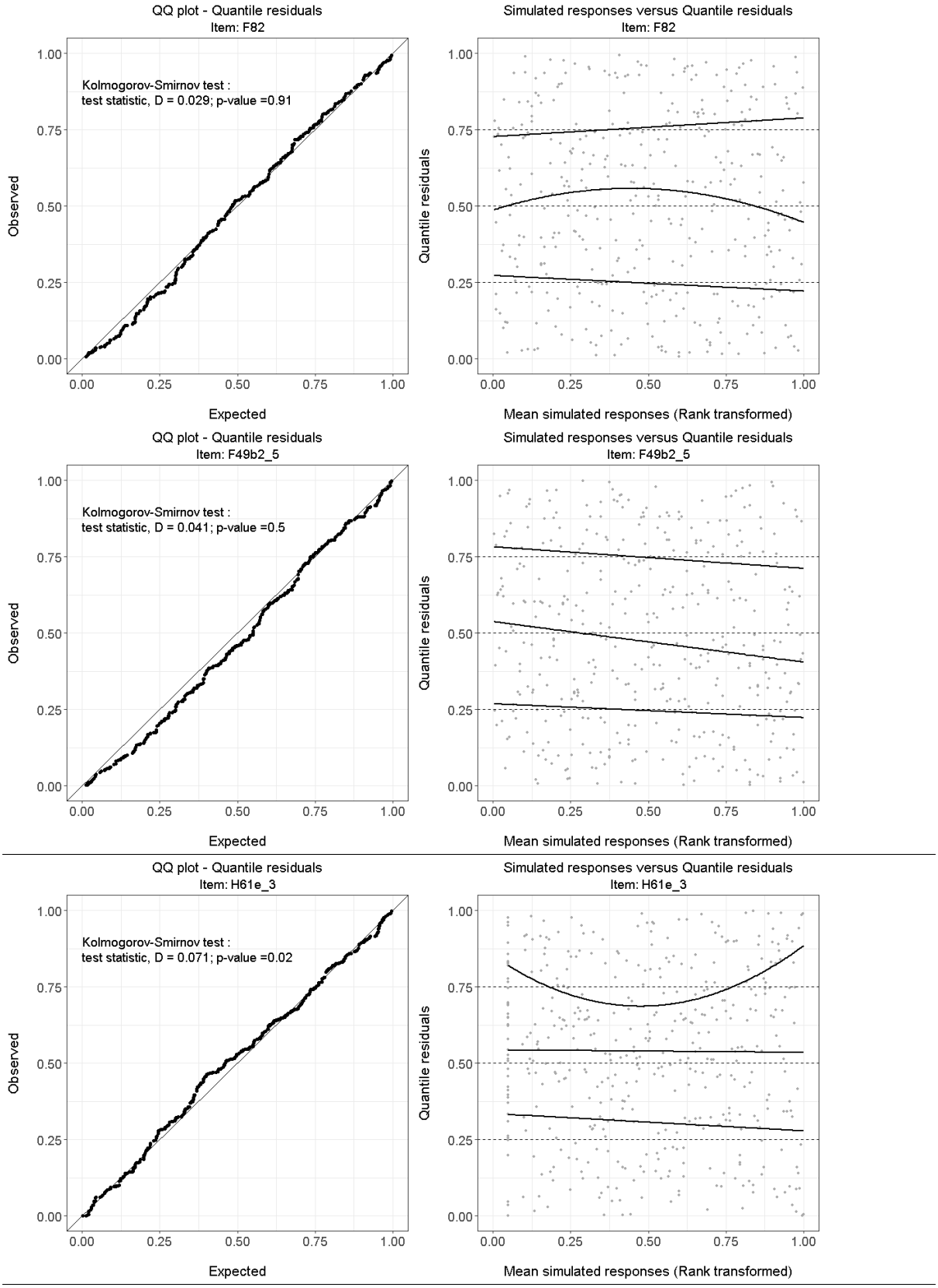

Figure 3.7 (continued).

3.6.3 Research question 3 - Do quality indicators explain interviewer non-response variance?

Our interest here lies in the explanation of interviewer non-response variance by the flag and non-flag variables. Table 3.9 shows the variance components for two outcomes - overall item non-response and non-response only due to DK. The variance components are all statistically significant and we see that the interviewer variance components are small compared to the item and respondent components. The covariance term in equation 3.9 was not significant and was dropped. The model with the non-response only due to 'NA/Refused' failed to be estimated by the software.

Table 3.9: Variance components for non-response models. The outcomes are the overall non-response and non-response only due to DK.

	Outcome variable: Overall item non-response (DK and NA/Refused)			Outcome variable: Non-response only due to DK		
	Item	Respondent	Interviewer	Item	Respondent	Interviewer
Base model	4.3	3.1	**0.18**	5.1	2.8	**0.20**
Non-flag model	4.5	3.1	**0.12**	5.2	2.8	**0.15**
Flag model	4.4	3.1	**0.09**	5.2	2.8	**0.12**
Full model	4.4	3.1	**0.08**	5.2	2.8	**0.11**

Figure 3.12 plots $\hat{p}_{ExplVar}$ for the non-flag, flag, and full models using the interviewer variance components from Table 3.9. We find 4 salient results. First, $\hat{p}^{Full}_{ExplVar}$ for either outcome is fairly substantial - 56% and 47% for the overall non-response and only-DK outcomes respectively; second, the flag model substantially outperforms the non-flag model - $\hat{p}_{ExplVar}$ of 52% versus 33% respectively for the overall non-response outcome and 42% versus 25% respectively for non-response only due to DK. These findings reflect previous research (e.g., Pickery and Loosveldt 1998) finding that typical interviewer-level variables (i.e., non-flag variables) fail to explain non-response interviewer variance; third, the full model does not add much over the flag model showing that, on average across items, the non-flag variables are not explaining any additional variance. This result is in contrast to the result for interviewer response error variance in research question 1 where the two sets of variables were complementary to each other; finally, $\hat{p}_{ExplVar}$ for the only-DK model are lower than the overall item non-response outcome model. This can occur if relatively more DKs are 'genuine' as compared to the 'NA/refusals', with respect to the protocols that the QC system are evaluating interviewers.

The odds ratios for the full model are plotted in Figure 3.13. Based on $\hat{p}_{ExplVar}$, flag models (and therefore also the full models) for the overall non-response outcome (left

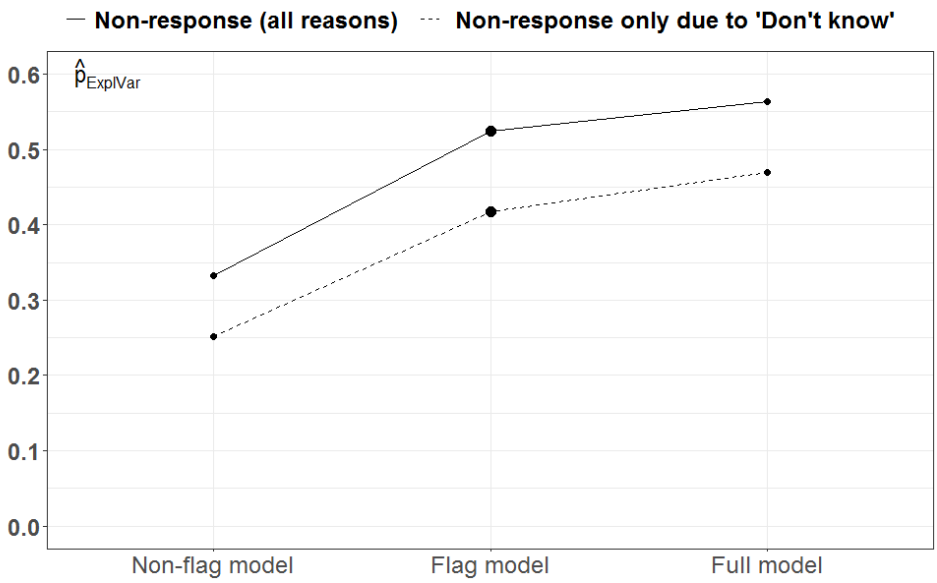

Figure 3.12: Proportions of variance explained by various non-response models. The $\hat{p}_{ExplVar}$ for non-flag, flag, and full models are plotted for two different outcomes: Overall non-response (DK and NA/Refused) and non-response only due to DK.

panel) used a quadratic transformation of *OverallFlagProportion* and a linear term of *ItemFlagProportion* while the only-DK outcome (right panel) used a quadratic transformation of both flag proportion variables. *ItemFlagProportion* is statistically significant in case of the only-DK outcome but not for the overall non-response outcome, while *OverallFlagProportion* is significant for both outcomes. This suggests that interviewers face item-specific challenges in eliciting responses when respondents finally give a DK answer while overall performance (i.e, averaged across items) impacts both DKs and NA/Refusals. Based on the effect sizes, *OverallFlagProportion* seems to be a more important interviewer-level variable than $ItemFlagProportion$ in predicting either outcome - this result too is in contrast to what we obtained for research question 1. The opposite direction of the quadratic term for *OverallFlagProportion* has an effect of slightly dampening the effects at the higher end of the proportions (the maximum overall flag proportion in the data is 0.1).

Figure 3.14 shows the diagnostic plot for the non-response (only 'don't know') model; we do not see any model fit issues. The plot for the non-response (all reasons) is similar and is therefore not shown.

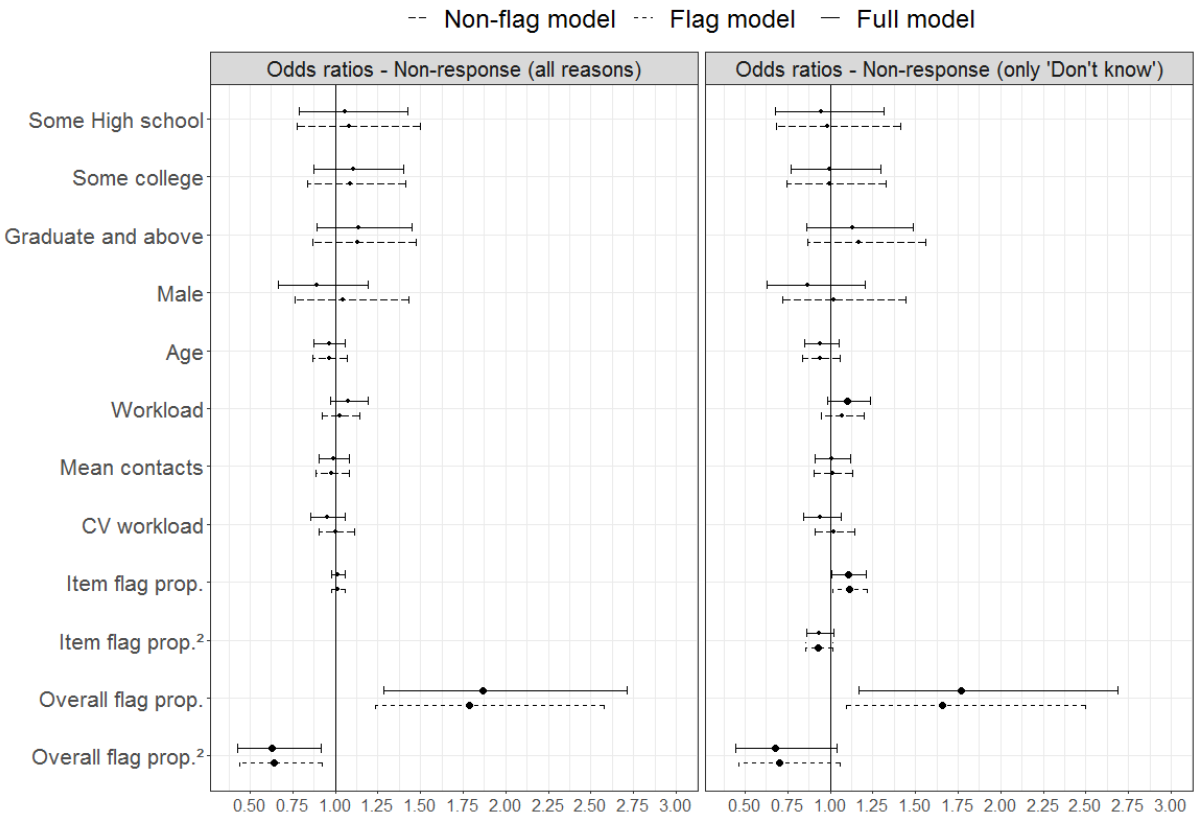

Figure 3.13: Odds ratios for overall non-response versus that due to only DK for the flag, non-flag, and full models. Results for the overall non-response outcome is plotted in the left panel and results for non-response only due to DK are plotted in the right panel. Horizontal bars are the 95% confidence intervals and bold points show those effects which are significant at a significance level of 0.1. The non-flag variables are all interviewer-level variables.

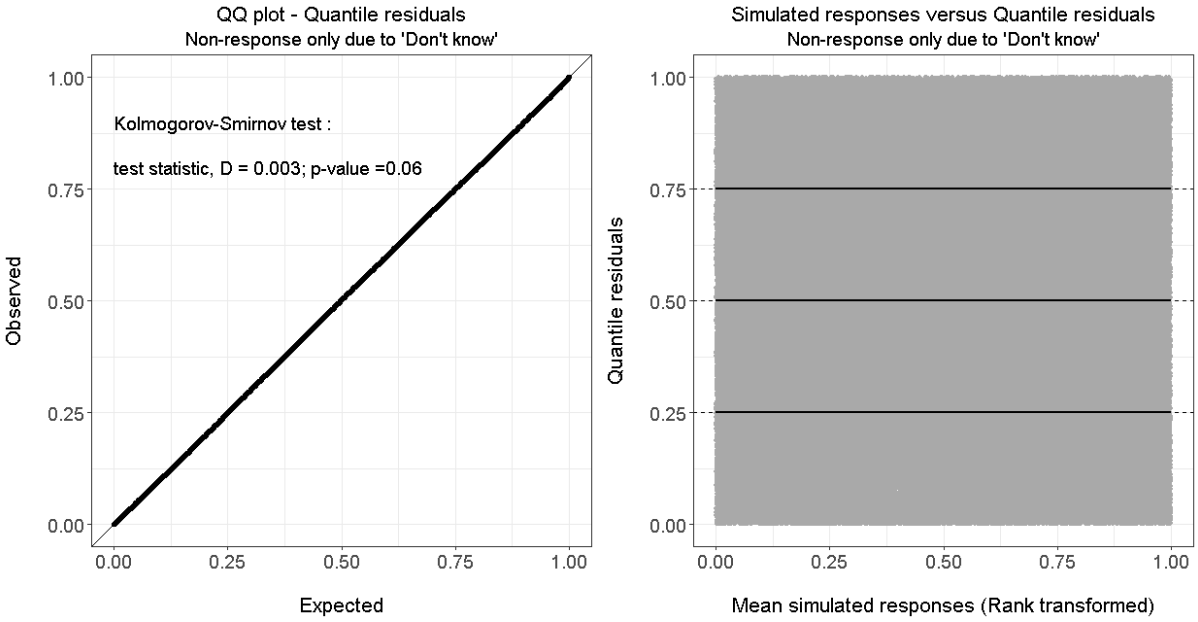

Figure 3.14: Diagnostic plots for the 'don't know' interviewer variance model. The left panel compares the quantile residuals to draws from a uniform distribution. Each point in the right panel is the mean simulated response (across 1000 simulations) for an observation in the data; the plot appears shaded in grey due to the large number of points. The solid lines in the correspond to the quantile regression lines and the dotted lines are benchmarks for these lines.

3.6.4 Research question 4 - Associations between item non-response and interviewing quality indicators

Here we explore associations between the case-level flag variables and item non-response. Before viewing the formal model results, we plot non-response rates within each QC flag category in Figure 3.15. We find that the two sub-categories of the major flag category have distinct non-response rates: 14% of the *ProbingFailure* cases are associated with a non-response compared to only 3% of the *OtherMajor* category. Non-response rates in the two major flag categories and the *Minor* flag category are much greater than the *NoFlag* category.

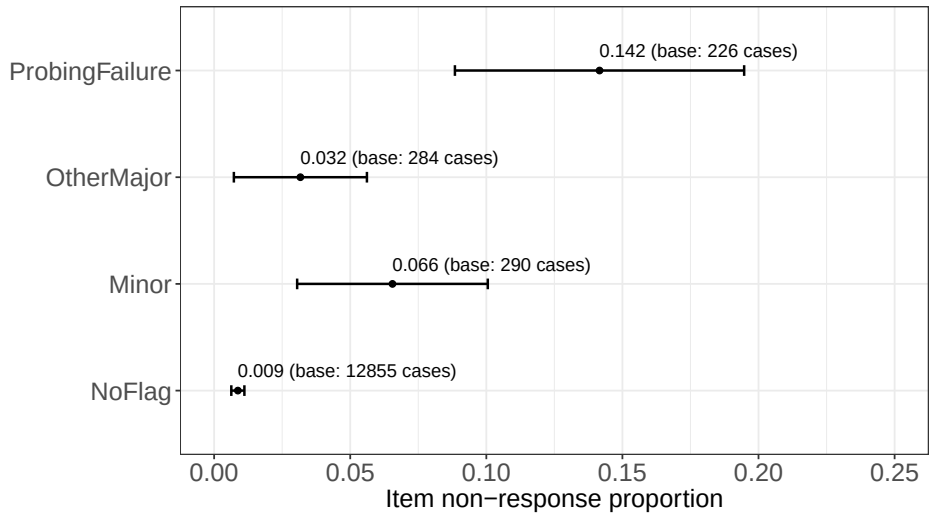

Figure 3.15: Non-response proportions in the four QC flag categories. The horizontal bars are the 95% confidence intervals that take into account the clustering of the (item-level) cases within interviewer.

Results of our formal model, which controls for respondent characteristics, are shown in Table 3.10 and reinforce the evidence seen above. Apart from statistically significant main effects for all flag terms, we find statistically significant interactions between *ProbingFailure* and *Mean daily workload*, and *Minor* and *some High school* interviewer education. The *OtherMajor* variable does not interact with any non-flag variable. Table 3.10 shows that among the random effects, only the item and interview-level random intercepts were significant but these are not our estimates of interest.

Interpretation of the coefficients in Table 3.10 is not straightforward given the presence of interactions. Figure 3.16 therefore plots predicted probabilities of non-response for the different flag variables, different levels of interviewer education, and different magnitudes of scaled interviewer mean daily workloads. The interviewers were assumed to be female, of average interviewer age, having an average workload, and average CV of daily workload. We used the following 'average' respondent to compute our predictions: a female high school graduate between 35 and 55 years of age who lives in 2 adult-1 child household

Table 3.10: Variance components and model coefficients for the case-level item non-response model. The interviewer variance component was small in magnitude and not significant, and is therefore not shown. Terms in bold are those which are significant at a 0.1 level.

Variance components
$\hat{\nu}^2_{0Resp} = 3.3$
$\hat{\nu}^2_{0Item} = 2.5$

	Log-odds	SE	p-value
Intercept	**-6.32**	**0.87**	**$\ll$ 0.001**
QC flag variables (implicit reference level: 'No flag')			
ProbingFailure	**3.00**	**0.32**	**$\ll$0.001**
OtherMajor	**1.26**	**0.42**	**0.002**
Minor	**2.04**	**0.60**	**0.001**
Interviewer-level covariates			
Male interviewer	0.08	0.44	0.93
Education (reference level: High School graduate)			
Some High school	-0.28	0.57	0.67
Some college	0.46	0.38	0.25
Graduate and above	**0.72**	**0.39**	**0.07**
Age (scaled)	-0.18	0.15	0.23
Workload (scaled)	0.03	0.16	0.86
Mean daily workload (scaled)	-0.25	0.17	0.15
CV of daily workload (scaled)	-0.13	0.14	0.41
Interaction terms: QC flag and Interviewer covariates			
Probing failure : Mean daily workload (scaled)	**1.08**	**0.31**	**$\ll$ 0.001**
Minor flag: Some High school	**2.07**	**0.95**	**0.03**
Minor flag: Some college	-0.48	0.86	0.62
Minor flag: Graduate and above	-1.47	0.94	0.12

with an annual income between $50,000 and $75,000.

We see that the predicted probabilities of non-response for cases with 'major flags for reasons other than probing' (third panel) are not different from the very small predicted probabilities for the *NoFlag* cases (first panel). This is true for the minor flag cases too except for those cases conducted by interviewers with the lowest education level with a relatively light average daily workload. The trend among the $ProbingFailure$ cases is in the opposite direction: when higher educated interviewers with a relatively higher mean daily workload fail to probe, those cases are associated with a greater non-response propensity. These probabilities represent a large increase in relative risk over the *NoFlag* cases.

Figure 3.17 shows the diagnostic plot for the model in Table 3.10; we do not see any model fit issues.

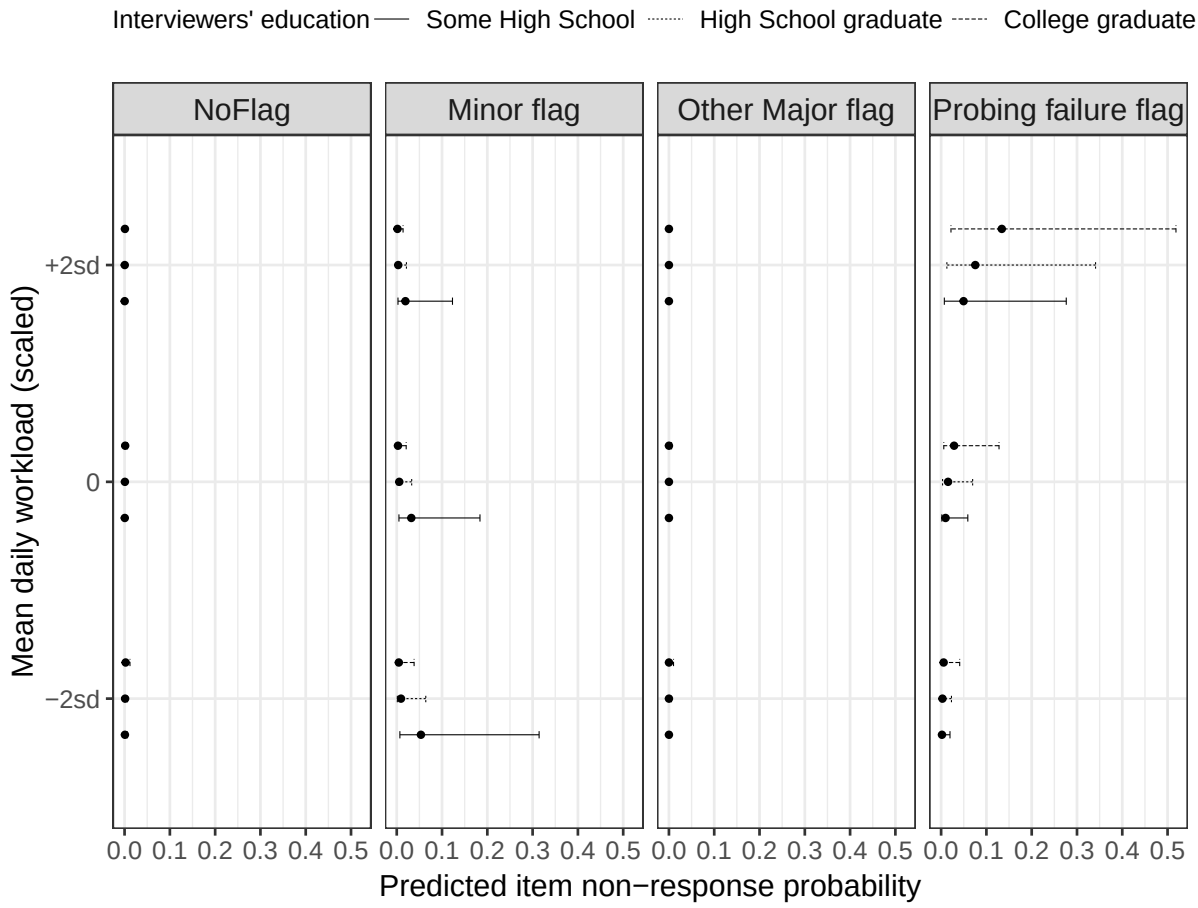

Figure 3.16: Plot of predicted item non-responses for different QC flags and interviewer covariates. The 4 different QC outcomes are shown in the 4 panels of the plot. The two interviewer level covariates are plotted - scaled mean daily workload (vertical axis, where 'sd' refers to standard deviation) and interviewer education (horizontal axis). The horizontal lines around each point are the 95% prediction intervals.

3.7 Discussion

Quality control systems seek to control interviewer effects by trying to detect interviewers' deviations from interviewing protocol. This research used data from a well-established QC system to investigate if detected deviations are actually associated with response and non-response measurement errors. We also checked if QC data have any incremental utility over traditional interviewer variables in this regard - this is important since QC processes can be expensive and time consuming.

Our results adduced reasonable evidence on both counts. In research question 1, we saw that while the QC flag variables were themselves able to only modestly explain interviewer effects, their complementarity with the non-flag variables led to a fairly substantial overall proportion of variance explained. Also, for 7 of the 39 models, more than a quarter of the interviewer variance was explained by the flag variables alone. This means that if

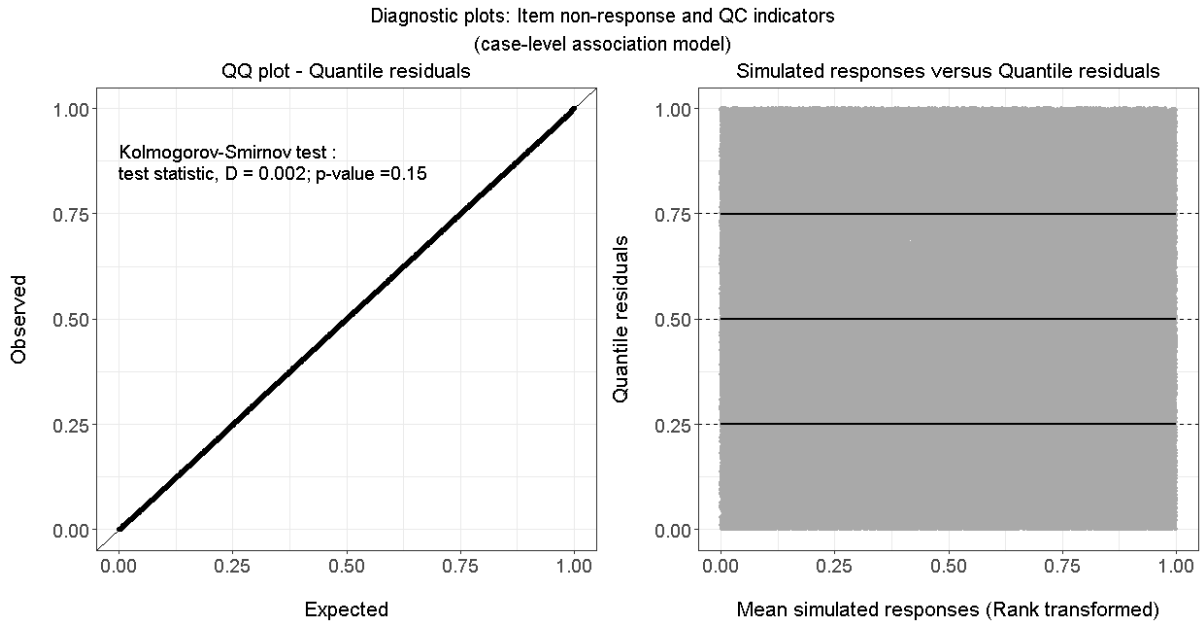

Figure 3.17: Diagnostic plots for the item non-response case-level associations model. The left panel compares the quantile residuals to draws from a uniform distribution. Each point in the right panel is the mean simulated response (across 1000 simulations) for an observation in the data; the plot appears shaded in grey due to the large number of points. The solid lines in the correspond to the quantile regression lines and the dotted lines are benchmarks for these lines.

the QC process is followed-up with quick and focused retraining (say, via audio/video calls) resulting in better interviewer behavior, we could theoretically expect a reduction in variance by a quarter for such items (for subsequently collected data) - a substantial reduction that can impact substantive inferences.

But what can explain why the interviewer effects are explained only moderately by the flag variables in research question 1? There are four possible reasons. First, the interviewer-level flag proportion variables are based on a small number of evaluations *per se* (median 6 evaluations per interviewer per item as pointed out in Section 3.4.2) which means that these predictors themselves have a lot of uncertainty. Second, the small number of evaluations also meant that we could not use specific behavioral indicators which would have been more effective as seen in research question 4. Third, the flag proportions are built out of indicator flag variables that perhaps do not have the granularity nor the flexibility to represent a complex phenomenon such as interviewer behavior. This may perhaps explain why proportions of variance explained by the flag variables for the continuous variables were not statistically significant. Fourth, the flag variables are based on a fixed set of behaviors based on interviewing protocols. It is possible that interviewer effects are associated with other finer dimensions such as paralinguistic aspects (Draisma and Dijkstra 2004) which are not captured in this QC process.

Based on these points, researchers wanting to study associations of quality control indicators with measurement error will benefit from designs that have a sufficient number of

evaluations per item that allow for robust and flexible analyses, even if this means having to reduce the total number of study items due to cost reasons. Also, rather than use a fixed set of behaviors, an open-ended approach of having coders evaluate all observed behaviors such as interviewer disfluencies, answer backchannels, and overspeech (Conrad et al. 2010) will yield richer data for analyses. Especially for telephone surveys where the interviewer can only establish presence via her voice, vocal characteristics might play a critical role, e.g., when the interviewer pauses or uses fillers such as "um", it could create a relaxed impression perhaps leading to more thoughtful responses by the respondent (Conrad et al. 2010; Christenfeld 1995). Results from this research can lead to reviewing the Quality Assurance (QA) process which lays down the interviewing protocols to be followed, by placing more emphasis on those behavioral aspects that impact interviewer effects the most.

Our construction of confidence intervals around $\hat{p}_{ExplVar}$ in research question 1 showed that it is important to compute measures of uncertainty for this statistic, something that does not seem to be commonly done. While the BCa is generally considered the best available procedure for constructing bootstrap confidence intervals (Efron and Tibshirani 1993, p.170; Carpenter and Bithell 2000; Puth et al. 2015) and our results appear reasonable, there are two cautionary notes. First, an examination of bias estimates and the acceleration constants (Appendix 3.C) as recommended by Efron and Hastie 2016 (p.195) shows that while the latter values were small, the former were generally large (Efron and Hastie (2016) suggest that these be less than or equal to 0.2). This means that the confidence intervals need to interpreted cautiously. Second, the definition of acceleration constant is 'not well-defined for multilevel models' and it is 'doubtful whether the jackknife for multilevel models will give a reasonable estimate of a third-order moment' (Leeden et al. 2008, p.419). We also computed percentile intervals (Efron and Tibshirani 1993, p. 170) but the BCa intervals, while producing fewer significant results than the percentile intervals, produced more reasonable results (percentile intervals spanned almost the entire range of possible values for many items). While computing bootstrap CIs for Research Question 3 was computationally prohibitive, a feasible approach might be to compute Jackknife (leave-one-interviewer-out) intervals. We hope that the production of confidence intervals for $\hat{p}_{ExplVar}$ in this research - to the best of our knowledge, the first in the survey methodology literature to do so - stimulates more research in this regard.

Turning to research question 2, the results point to possible bias for the few items where we found significant fixed effects. But given that only a tiny proportion of cases were found to be involved and that we are working with early release data (and not final data where these anomalies would most likely be corrected by methods such as imputation), the bias levels would be small; yet, the results are important in shedding light on properties of the QC flag variables. Except for item F49b2_5 (not a hybrid, electric and alternate

fuel vehicle type), none of the items in Tables 3.7 and 3.8 had a significant relationship in research question 1 thereby demonstrating the value of conducting a case-level analysis as done here. While the results were intuitively appealing, they should be interpreted cautiously since they could be causally driven by common cause variables other than interviewer behaviors. Also, the reasoning that we offered for the results can only be verified by listening to the recordings but we unfortunately did not have access to these.

As compared to explaining interviewer response variance (research question 1), the QC flag variables were more successful in explaining non-response variance (research question 3). A skeptical explanation is that the outcome is very visible in the case of non-response; since QC coders for this project were experienced with many of them also having interviewing experience, they would be sensitive to the fact that non-response is undesirable, making them prone to flag incidences of non-response. However, we do not see indiscriminate flagging of non-response cases; of the 172 evaluated non-response cases for the 27 analyzed items, interviewing in 112 cases (65%) were actually evaluated as being positive. The results show that the QC process is successful at picking up behaviors that predict non-response. Reading results from research questions 2 and 4 together suggests that lack of probing is an important correlate of both response and non-response interviewer variance, a finding in line with previous research (e.g., Fowler and Mangione 1990).

Results of our research will be conservative due to three reasons. First, interviewers are aware of which interviews are recorded and they tend to adhere to better interviewing standards for these (McGonagle et al. 2015). Second, our evaluation of interviewer effects is based on the data finally recorded by the interviewer but at least some values would have undergone edits during the survey. This can happen when the interviewer realizes an inconsistency in a later question and returns to edit the response, or when the computer throws up an error message forcing the interviewer to edit the response. In such cases, the value gets corrected but the case would still be flagged if it was an interviewing behavior that led to the initial error. This reduces the magnitude of the associations between the interviewing quality indicators and interviewer effects. Third, we did not include interactions between the flag and non-flag covariates in research questions 1 and 3 since we were primarily interested in checking the incremental value of the flag variables over the non-flag variables in explaining variance; including interactions would potentially increase the proportion of variance explained by the full model.

We mention seven limitations of this study. First, we will not know if we are successful in making a good approximation to a true interpenetrated design. If there are situations where difficult respondents (where difficulty cannot be accounted for by $\boldsymbol{X}$) are alloted to interviewers only of a certain profile, respondent effects will masquerade as interviewer effects. We regard this as an area of future research. Second, our analysis assumes

that the QC coding itself has no measurement error. Unfortunately, no inter-reliability scores were available to assess this. Third, research questions 1, 3, and 4 focused on overall patterns of associations between QC variables and interviewer effects and we did not delve into item-specific explanations. Future research can consider questions such as which item types are better explained by flag (or non-flag) variables and why, e.g., we found that interviewer variance for none of the numeric variables were explained by the flag variables in research question 1. Fourth, our analysis ignores the panel aspect of the study survey. A potentially important non-flag interviewer variable could have been the proportion of the current workload defined by the same respondents as the previous wave; familiarity may help the respondent be more forthcoming with responses but it could also lead the interviewer to be less careful with interviewing. Fifth, effects for items subject to dependent interviewing (which arises given the panel nature of the study) could be different from the rest (Pascale and McGee 2008) but we did not access information on which items were pre-filled for the interviewer. Sixth, West and Olson (2010) show that there is a component of interviewer variance that is also due to non-response error variance and not measurement error variance alone. This may not matter in the case of PSID with a very high wave-on-wave response rate, but methods to be able to parse out the measurement error component may be important for other surveys. Finally, due to reasons explained earlier, our analysis did not include items in the EHC which is a core part of the PSID questionnaire and which interviewers find complex (Mcgonagle 2013). Our recommendation is to code specific items in the EHC so these can be analyzed in the future.

3.8 Implications for survey practice and conclusion

The complementarity of the flag and non-flag variables that we found in answering research question 1 is an opportunity to increase efficiency of the QC process. Figure 3.18 shows one possible strategy based on the proportion of variance explained by the flag variables (vertical axis) and its absolute difference with the non-flag variables (horizontal axis).

Those items where the flag variables explain a high proportion of variance performing much better than the non-flag variables (quadrant 1) should get the highest emphasis in the QC evaluation process. On the other extreme is quadrant 3 where flag and non-flag variables do not explain interviewer effects. More research is needed on the exploration of the possible mechanisms that generate interviewer effects for these items. Quadrants 2 and 4 are items where the non-flag variables do better than (quadrant 2) or somewhat equal to (quadrant 4) the flag variables. Here, one can use the non-flag model of a

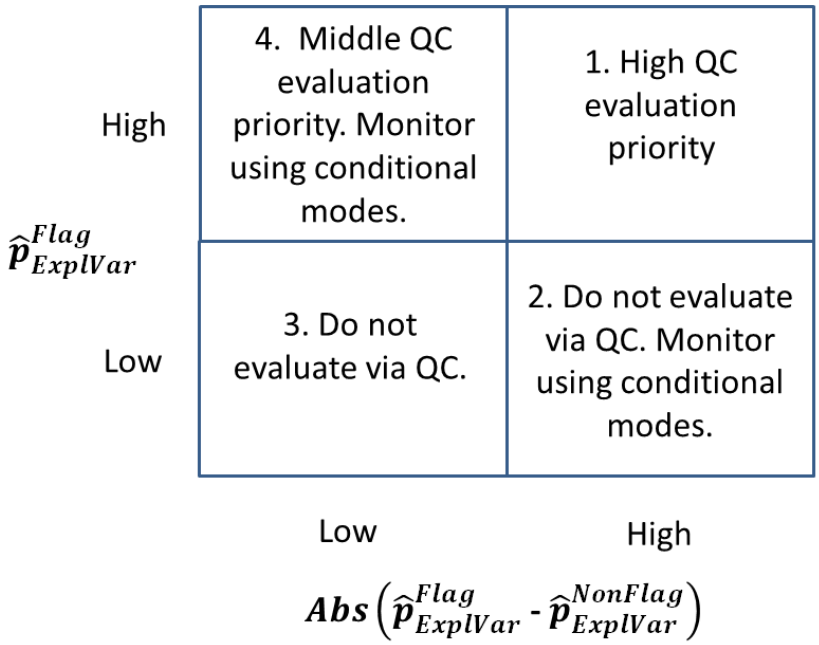

Figure 3.18: Prioritization of items in a QC process

preceding survey wave to predict the current wave's conditional modes. Interviewers with the high magnitudes of predicted conditional modes can then be selected so as to listen to their recordings.

In answering research questions 1 and 3, we found that the overall flag variable was important in predicting item-specific interviewer effects. This means that refresher interviewer training should not only focus on administering specific items but also general interviewing behavior. Based on results in research question 4, QC systems should pay closer attention to interviewers with a high workload in order to reduce non-response.

Survey organizations often classify errors between major and minor (e.g., Couper et al. 1992; Mudryk et al. 1996) so that supervisors can prioritize cases when giving feedback to interviewers. While such classifications are useful, results for research questions 2 and 4 showed that the minor flag was also important in detecting interviewer effects/predicting non-response. This means that behaviors that coders might perceive as 'minor' can actually have a substantial impact on data quality (Marquis et al. 1972; Schuman and Presser 1981) thereby cautioning survey managers from ignoring minor flags. The exact behaviors that cause the minor flags should also be recorded.

Mudryk et al. (1996) list 7 characteristics of CATI quality control but none of these are directly linked to measurement error. We feel it is imperative to link quality control to survey error explicitly. Groves (1989, p.389) mentions that research that links such interviewing evaluations and measurement error "should receive highest priority in the future". This research is one step in this direction. Replications of this research on other

surveys, including face-to-face surveys, will serve to add to the evidence on hand.

The limitations of the QC data motivate the next chapter: as compared to these data, paradata have a wide range of variables, have information on all respondents, the flexibility to be transformed and interactions created among variables, are a better proxy for the interviewer-respondent interaction (evaluation data takes into account only the interviewer), and are objective (no reliance on human coders). It is worth exploring how paradata can be used to spot interviewers who may be injecting measurement error into estimates.

Appendix

3.A Summary descriptive statistics for the items used for the analyses.

Table 3.A.1: Descriptive statistics for the 27 variables used for the analyses. These are five-number summaries for the 14 numeric response items, and proportions for the 7 binary response variables. Items are arranged according to their order in the questionnaire. Item descriptions are not precise; please see the questionnaire for the detailed question.

Numeric response variables

Item	Item description	Minimum	Q1	Median	Mean	Q3	Maximum
A8	Number of rooms	0	4	5	5	7	20
A20	Present home value ($)	1	97,000	170,000	231,408	280,000	8,500,000
A21	Total yearly property tax ($)	1	900	1,900	2,865	3,596	4,000
A22	Total yearly homeowner's insurance ($)	1	600	950	1,102	1,300	9,000
A42A	Monthly gas expenses ($)	0	0	40	97	100	698
A42B	Monthly electricity expenses ($)	0	60	100	134	170	398
A43	Monthly water and Sewer expenses ($)	0	0	30	51	68	392
A44	Telephone and internet expenses ($)	0	90	175	196	270	436
F77	Monthly car insurance amount ($)	1	118	200	568	700	917
F80B	Monthly car gasoline expenses ($)	1	100	150	197	250	2,000
F81A	Monthly bus and train fare expenses ($)	0	0	0	9	0	100
F81B	Monthly taxicab expenses ($)	0	0	0	4	0	58
G13	Head's annual gross wages/salaries ($)	15	20,000	37,000	50,624	62,000	5,000,000
H61J	Monthly health insurance amount ($)	0	75	188	252	350	4,992

Binary response variables

F47	Any vehicle owned/leased for personal use? Yes = 84%
F57	Vehicle also used for business purposes? Yes = 17%
F82	Any school-related expenses in 2014? Yes = 24%
G12	Any salaries or wages besides uninc. business? Yes = 76%
G14	Any income from bonuses, overtime, tips, or commissions? Yes = 15%
H61K	Any family member without health insurance? Yes = 23%
KL84	Spouse attending or enrolled in regular school? Yes = 2%

Category proportions for the 6 multinomial response items. Proportions of categories that are analyzed are in bold. Categories are ordered as they appear in the questionnaire. Item H61E does not include 'Other state sponsored plan', 'Indian health insurance', and 'Other government plan' for which the proportions are close to 0. Item descriptions are not precise; please see the questionnaire for the detailed question.

A4 (Dwelling-unit type)	One-family house	Two-family house or duplex	Apartment in a multi-unit	Mobile home or trailer	Row or town house	Other-Specify		
	65%	**4%**	**23%**	**5%**	2%	1%		
A41_1 (Home heating type)	Gas	Electricity	Oil	Wood	Coal	Solar	Bottled gas; propane	Kerosene	Other-specify
	51%	**40%**	**4%**	1%	0%	0%	2%	0%	1%
A42 (Utility bill type)	One Electricity/Gas/Fuel utility bill and can report separate amounts		One Electricity/Gas/Fuel utility bill but cannot report separate amounts		Electricity/Gas/Fuel utilities not on one bill			
	12%		**17%**		**71%**			
BCDE1 (Person 1 employment status)	Working	Temporarily laid off	Looking for work, unemployed	Retired	Permanently disabled	Housewife; keeping house	Student	
	67%	0.5%	**7%**	**12%**	**4%**	**7%**	1%	
F49B2 (Vehicle type)	Hybrid vehicle	Plug-in hybrid vehicle	Electric vehicle or battery	Alternative fuel vehicle (CNG, lpg)	None of the above	Other-Specify		
	2%	0%	0%	2%	**96%**	0%		
H61E (Person 1 type of health insurance)	Employer-provided health insurance	Private health insurance	Medicare	Medi-Gap/Supplemental	Medicaid/ State medical program	Military health care	Other health insurance	
	58%	**9%**	**15%**	0.2%	**14%**	**3%**	**0.4%**	

3.B R code for bias-corrected and accelerated confidence intervals

```
# boot_results is a dataframe containing the resampled statistics

# The estimate of interest in this example is the proportion of
# variance explained by the flag variables ('p_varexpl_flag');
# the vector of estimates in 'boot_results' is called
# 'p_varexpl_flag_boot' while the original sample estimate is
# called 'p_varexpl_flag_original'.

# z0 is an estimate of the bias
z0 <- qnorm((sum(boot_results[, p_varexpl_flag_boot] <
                                    p_varexpl_flag_original)/1000))

# standard normal quantiles corresponding to a 95% interval
z <- qnorm(c(0.05/2, 1-0.05/2))

# The 'acceleration_constant' comes from a Jackknife procedure
numer <- z0 + z
denom <- 1 - acceleration_constant*(z0 + z)

p = pnorm(z0 + numer/denom)

BCa_CIs <- quantile(boot_results[ , p_varexpl_flag_boot] ,
p=p, names=FALSE, na.rm=TRUE)
```

3.C Magnitudes of bootstrap bias corrections and acceleration constants

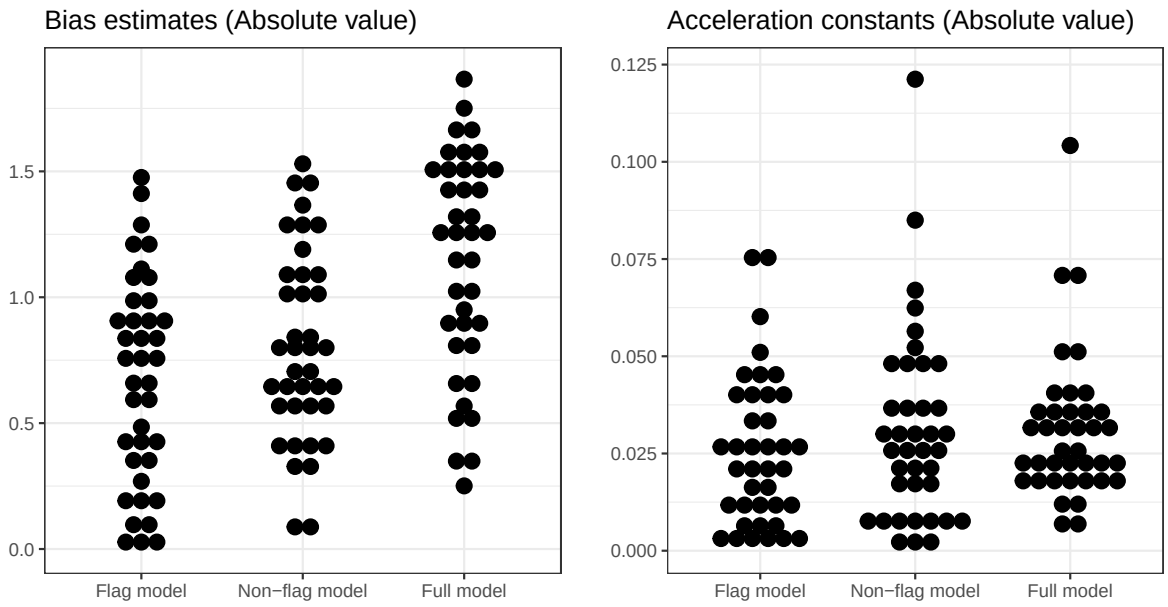

Figure 3.C.1: Bias estimates and acceleration constants for the BCa intervals. The bias estimates are plotted in the left panel and acceleration constants in the right panel for all three models. Each point in the plots is an item for which the models were fit. While the acceleration constants are fairly small we see evidence of some fairly large biases.

.

3.D R code for model diagnostics

```
require(DHARMa)
require(qrnn)
require(ggplot2)

#simulate observations
gof_model <- simulateResiduals(model, #name of the model
                               #1000 responses simulated
                               n = 1000,
                               refit = F,
                               #condition on all random effects
                               re.form = NULL)

#mean of the simulated responses for each observation
mean_simresponse <- gof_model$fittedPredictedResponse

#rank transform the mean simulated responses for better visualization
mean_simresponse <- rank(mean_simresponse, ties.method = "average")
mean_simresponse <- mean_simresponse/max(mean_simresponse)

#extract quantile residuals
scaled_resids <- gof_model$scaledResiduals

#data frame for plots
quantresids_data <- data.frame(scaled_resids = scaled_resids, #quantile
                               Expected = runif(gof_model$nObs),
                               mean_simresponse = mean_simresponse)

### Quantile regression
#penalty factor kept as 1 to reduce overfitting
#25th percentile
fit25_nl <- qrnn.fit(x = as.matrix(quantresids_data$mean_simresponse),
                     y = as.matrix(quantresids_data$scaled_resids),
                     n.hidden = 4, iter.max = 1000,
                     n.trials = 1, penalty = 1,
                     tau = 0.25)
quantresids_data$fit25_nl <- qrnn.predict(
  as.matrix(sort(quantresids_data$mean_simresponse)), fit25_nl)

#median
fit50_nl <- qrnn.fit(x = as.matrix(quantresids_data$mean_simresponse),
                     y = as.matrix(quantresids_data$scaled_resids),
                     n.hidden = 4, iter.max = 1000,
                     n.trials = 1, penalty = 1,
                     tau = 0.5)
```

```
quantresids_data$fit50_nl <- qrnn.predict(
  as.matrix(sort(quantresids_data$mean_simresponse)), fit50_nl)

#75th percentile
fit75_nl <- qrnn.fit(x = as.matrix(quantresids_data$mean_simresponse),
                     y = as.matrix(quantresids_data$scaled_resids),
                     n.hidden = 4, iter.max = 1000,
                     n.trials = 1, penalty = 1,
                     tau = 0.75)
quantresids_data$fit75_nl <- qrnn.predict(
  as.matrix(sort(quantresids_data$mean_simresponse)), fit75_nl)

#Kolmogorov-Smirnov test - uniform reference distribution
#used for annotation in the QQ plot
ks.test(quantresids_data$scaled_resids, 'punif')


#QQ plot (theme elements and annotations not shown for brevity)
p_quantresids <- ggplot(quantresids_data,
                        aes(x = sort(Expected),
                            y = sort(scaled_resids))) +
  geom_abline(slope = 1, intercept = 0) +
  ggtitle("QQ plot - Quantile residuals",
  subtitle = "(Interview-level mode)1") +
  xlab("Expected") + ylab("Observed")

#Plot of mean simulated responses against quantile residuals
#(theme elements and annotations not shown for brevity)
p_fitted_quantresids <- ggplot(quantresids_data,
                               aes(x = mean_simresponse,
                               y = scaled_resids)) +

  #quantile regression lines
  geom_line(aes(x = sort(mean_simresponse), y = fit25_nl), size = 1) +
  geom_line(aes(x = sort(mean_simresponse), y = fit50_nl), size = 1) +
  geom_line(aes(x = sort(mean_simresponse), y = fit75_nl), size = 1) +

  #reference lines
  geom_abline(slope = 0, intercept = 0.25, linetype = 2) +
  geom_abline(slope = 0, intercept = 0.50, linetype = 2) +
  geom_abline(slope = 0, intercept = 0.75, linetype = 2) +

  ggtitle("Simulated responses versus Quantile residuals",
  subtitle = "(Interview-level model)") +
  xlab("Mean simulated responses (Rank transformed)") +
  ylab("Quantile residuals")
```

References

Andreski, P., Mcgonagle, K., and Schoeni, R. (2007). An Analysis of the Quality of the Health Data in the Panel Study of Income Dynamics - Technical Series Paper #07-04. Technical report, Institute for Social Research, University of Michigan, Ann Arbor.

Auriat, N. (1993). "My Wife Knows Best": A Comparison of Event Dating Accuracy Between the Wife, the Husband, the Couple, and the Belgium Population Register. *Public Opin. Q.*, 57(2):165.

Austin, P. C. and Merlo, J. (2017). Intermediate and advanced topics in multilevel logistic regression analysis. *Stat. Med.*, 36(20):3257–3277.

Bailar, B., Bailey, L., and Stevens, J. (1977). Measures of Interviewer Bias and Variance. *J. Mark. Res.*, 14(3):337.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.*, 67(1).

Beullens, K. and Loosveldt, G. (2016). Interviewer Effects in the European Social Survey. *Surv. Res. Methods*, 10(2):103–118.

Biemer, P., Herget, D., Morton, J., and Willis, G. (2000). The Feasibility of Monitoring Field Interview Performance Using Computer Audio-recorded Interviewing (CARI). In *Proc. Jt. Stat. Meet. Surv. Res. Methods Sect.*, pages 1068–1073. American Statistical Association.

Blair, E. (1980). Using Practice Interviews to Predict Interviewer Behaviors. *Public Opin. Q.*, 44(2):257.

Cannell, C., Fowler, F., and Marquis, K. (1968). The Influence of Interviewer and Respondent Psychological and Behavioral Variables on the Reporting of Household Interviews. *Vital Heal. Stat.*, Series 2(No. 26.).

Cannell, C., Lawson, S., and Hausser, D. (1975). A Technique for Evaluating Interviewer Performance. Technical report, Institute for Social Research, The University of Michigan, Ann Arbor.

Cannell, C., Miller, P., and Oksenberg, L. (1981). Research on interviewing techniques. *Sociol. Methodol.*, 12:389–437.

Cannon, A. (2011). Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Comput. Geosci.*, 37(9):1277–1284.

Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat. Med.*, 19(9):1141–1164.

Chapman, D. and Weinstein, R. (1988). Design of a Systematic Monitoring Plan for CATI Surveys. In *JSM Proceedings, Surv. Res. Methods Sect.*, pages 543–548. American Statistical Association.

Christenfeld, N. (1995). Does It Hurt to Say Um? *J. Nonverbal Behav.*, 19(3):171–186.

Collins, W. (1970). Interviewers' Verbal Idiosyncrasies as a Source of Bias. *Public Opin. Q.*, 34(3):416.

Conrad, F., Broome, J., Benki, J., Groves, R., Kreuter, F., and Vannette, D. (2010). To Agree or Not to Agree? Impact of interviewer speech on survey participation decisions. In *Am. Stat. Assoc. Proc. Sect. Surv. Res. Methods*, pages 5979–5993, Vancouver, Canada.

Couper, M., Holland, L., and Groves, R. (1992). Developing systematic procedures for monitoring in a centralized telephone facility. *J. Off. Stat.*, 8(1):63–76.

Currivan, D., Dean, E., and Thalji, L. (2006). Using Standardized Interviewing Principles to Improve a Telephone Interviewer Monitoring Protocol. In *2nd Int. Conf. Teleph. Surv. Methodol.*, Miami Beach, FL.

Davis, R., Couper, M., Janz, N., Caldwell, C., and Resnicow, K. (2010). Interviewer effects in public health surveys. *Health Educ. Res.*, 25(1):14–26.

Draisma, S. and Dijkstra, W. (2004). Response Latency and (Para)Linguistic Expressions as Indicators of Response Error. In PresserS., Rothgeb, J., Couper, M., Lessler, J., Martin, E., Martin, J., and Singer, E., editors, *Methods Test. Eval. Surv. Quest.*, pages 131–147. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Dunn, P. and Smyth, G. (1996). Randomized Quantile Residuals. *J. Comput. Graph. Stat.*, 5(3):236.

Efron, B. (1987). Better Bootstrap Confidence Intervals. *J. Am. Stat. Assoc.*, 82(397):171–185.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Springer US, Boston, MA.

Elliott, M. and West, B. (2015). Clustering by Interviewer: A Source of Variance That Is Unaccounted for in Single-Stage Health Surveys. *Am. J. Epidemiol.*, 182(2):118–126.

Fowler, F. and Mangione, T. (1990). *Standardized Survey Interviewing: Minimizing Interviewer Related Error*. SAGE Publications, Inc., Thousand Oaks, California.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Groves, R. (1989). *Survey Errors and Survey Costs*. John Wiley & Sons, Inc., Hoboken,

NJ, USA.

Groves, R. and Magilavy, L. (1986). Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys. *Public Opin. Q.*, 50(2):251.

Hartig, F. (2018). DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. `https://cran.r-project.org/package=DHARMa`.

Hicks, W., Edwards, B., Tourangeau, K., McBride, B., Harris-Kojetin, L., and Moss, A. (2010). Using Cari Tools To Understand Measurement Error. *Public Opin. Q.*, 74(5):985–1003.

Hildum, D. C. and Brown, R. W. (1956). Verbal reinforcement and interviewer bias. *J. Abnorm. Soc. Psychol.*, 53(1):108–111.

Hox, J. (1994). Hierarchical Regression Models for Interviewer and Respondent Effects. *Sociol. Methods Res.*, 22(3):300–318.

Hox, J. (2010). *Multilevel Analysis: Techniques and Applications*. Routledge, New York, NY, second edition.

Hyman, H. (1954). *Interviewing in Social Research*. University of Chicago Press, Chicago, IL.

Knauper, B. (1999). The Impact of Age and Education on Response Order Effects in Attitude Measurement. *Public Opin. Q.*, 63(3):347.

Krosnick, J. and Alwin, D. (1987). An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opin. Q.*, 51(2):201.

Lee, J. and Lee, S. (2012). Does it Matter WHO Responded to the Survey? Trends in the U.S. Gender Earnings Gap Revisited. *ILR Rev.*, 65(1):148–160.

Leeden, R., Meijer, E., and Busing, F. (2008). Resampling Multilevel Models. In *Handb. Multilevel Anal.*, pages 401–433. Springer New York, New York, NY.

Loosveldt, G. and Beullens, K. (2014). A Procedure to Assess Interviewer Effects on Nonresponse Bias. *SAGE Open*, 4(1):1–12.

Mahalanobis, P. (1946). Recent Experiments in Statistical Sampling in the Indian Statistical Institute. *J. R. Stat. Soc.*, 109:325–370.

Mangione, T., Fowler, F., and Louis, T. (1992). Question Characteristics and Interviewer Effects. *J. Off. Stat.*, 8:293–307.

Marquis, K. (1969). Interviewer-Respondent interaction in a household interview. In *Proc. Jt. Stat. Meet. Surv. Res. Methods Sect.* American Statistical Association.

Marquis, K., Cannell, C., and Laurent, A. (1972). Reporting health events in household interviews: effects of reinforcement, question length, and reinterviews. *Vital Heal. Stat. Natl. Cent. Heal. Stat.*, 2(45).

Mathiowetz, N. and Cannell, C. (1980). Coding Interviewer behavior as a method of evaluating performance. In *Proc. Jt. Stat. Meet. Surv. Res. Methods Sect.*, pages 525–

528, Houston, TX. American Statistical Association.

Mcgonagle, K. (2013). Survey Breakoffs in a Computer-Assisted Telephone Interview. *Surv. Res. Methods, J. Eur. Surv. Res. Assoc.*, 7(2):79–90.

McGonagle, K., Brown, C., and Schoeni, R. (2015). The Effects of Respondents' Consent to Be Recorded on Interview Length and Data Quality in a National Panel Study. *Field methods*, 27(4):373–390.

Mitchell, S., Strobl, M., Fahrney, M., Nguyen, T., Bibb, S., Thissen, R., and Stephenson, W. (2008). Using computer audio-recorded interviewing to assess interviewer coding error. In *Proc. Proc. Jt. Stat. Meet. Surv. Res. Methods Sect.*, Alexandria, VA. American Statistical Association.

Mudryk, W., Burgess, M., and Xiao, P. (1996). Quality control of CATI operations in Statistics Canada. In *Proc. Jt. Stat. Meet. Sect. Surv. Res. Methods*, pages 150–159.

Ongena, Y. and Dijkstra, W. (2006). Methods of Behavior Coding of Survey Interviews. *J. Off. Stat.*, 22(3):419–451.

Pascale, J. and McGee, A. (2008). Using behavior coding to evaluate the effectiveness of dependent interviewing. *Surv. Methodol.*, 34(2):143–151.

Pickery, J. and Loosveldt, G. (1998). The Impact of Respondent and Interviewer Characteristics on the Number of No Opinion Answers. *Qual. Quant.*, 32(1).

Powell, M. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. Technical report, Department of Applied Mathematics and Theoretical Physics, Cambridge, England.

Puth, M.-T., Neuhäuser, M., and Ruxton, G. (2015). On the variety of methods for calculating confidence intervals by bootstrapping. *J. Anim. Ecol.*, 84(4):892–897.

Rustemeyer, A. (1977). Interviewer Performance in Mock Interviews. In *Proc. Jt. Stat. Meet. Soc. Stat. Sect.*, pages 341–346. American Statistical Association.

Schuman, H. and Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Academic Press, New York.

Self, S. and Liang, K.-Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. *J. Am. Stat. Assoc.*, 82(398):605–610.

Shoemaker, P., Eichholz, M., and Skewes, E. (2002). Item Nonresponse: Distinguishing between don't Know and Refuse. *Int. J. Public Opin. Res.*, 14(2):193–201.

Skowronski, J. and Thompson, C. (1990). Reconstructing the dates of personal events: Gender differences in accuracy. *Appl. Cogn. Psychol.*, 4(5):371–381.

Snijders, T. and Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage Publishers, London, 1st edition.

Sudman, S., Bradburn, N., Blair, E., and Stocking, C. (1977). Modest Expectations: The Effects of Interviewers' Prior Expectations on Responses. *Sociol. Methods Res.*,

6(2):171–182.

Tarnai, J. (2007). Monitoring CATI Interviewers. In *62nd Annu. Conf. Am. Assoc. Public Opin. Res.*, Anaheim, California.

Tarnai, J. and Moore, D. (2007). Measuring and Improving Telephone Interviewer Performance and Productivity. In Lepkowski, J., Tucker, C., M., B., de Leeuw, E., Japec, L., Lavrakas, P., Link, M., and Sangster, R., editors, *Adv. Teleph. Surv. Methodol.*, pages 359–384. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Team, R. C. (2013). R: A Language and Environment for Statistical Computing.

Thissen, R. (2014). Computer Audio-Recorded Interviewing as a Tool for Survey Research. *Soc. Sci. Comput. Rev.*, 32(1):90–104.

Thissen, R., Sattaluri, S., McFarlane, E., and Biemer, P. (2008). The Evolution of Audio Recording in Field Surveys. *Surv. Pract.*, 1(5).

West, B., Kreuter, F., and Jaenichen, U. (2013). Interviewer Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error, or Nonresponse? *J. Off. Stat.*, 29(2).

West, B. and Olson, K. (2010). How Much of Interviewer Variance is Really Nonresponse Error Variance? *Public Opin. Q.*, 74(5):1004–1026.

Ypma, J., Borchers, H., and Eddelbuettel, D. (2014). nloptr - R interface to NLopt.

Chapter 4

Can paradata predict interviewer effects?

4.1 Introduction

4.1.1 Interviewer effects

The positive role of interviewers in surveys has long been recognized; interviewers solicit cooperation, motivate and probe respondents to provide accurate responses, assist in clarifying questions, and record responses (Groves 1989, p.359; Groves and Couper 1998, Chapter 7; Groves et al. 2004, p.141). On the other hand, interviewers are also a source of error. Ideally, interviewers in a survey should be exchangeable. Holding all other survey conditions constant, we should obtain the same response from a respondent irrespective of which interviewer undertakes the interview. But interviewers vary in their behaviors thereby impacting responses. Even if two different interviewers read questions exactly as worded in the instrument, they might pace their delivery differently thus affecting respondent attention (Cannell et al. 1981). This results in answers from respondents interviewed by a particular interviewer tending to be similar to each other as compared to responses obtained from respondents belonging to other interviewers, even after factoring out differences in geography and respondent profile (Hansen et al. 1960; Fellegi 1974; Biemer and Stokes 1985). Thus, the expected response means across interviewers differ, even under an interpenetrated sample design (Mahalanobis 1946) with a 100% response rate. We use the term 'interviewer effects' to refer to this 'interviewer measurement error (response) variance'. The associated intra-interviewer correlation (ρ_{int}) is given by:

$$\rho_{int} = \frac{between\text{-}interviewer\ variance}{between\text{-}interviewer\ variance + within\text{-}interviewer\ variance} \tag{4.1}$$

Values of ρ_{int} are small - typically less than 0.02 (Groves 1989, p.318). However, the inflation in variance for a mean due to interviewer measurement error is given by $1 + (MeanWorkload - 1)\rho_{int}$. This means that if an item had a ρ_{int} of 0.02 when the mean interviewer workload is 50, the variance almost doubles, greatly reducing the precision of the estimate. This necessitates methods to control interviewer effects.

4.1.2 The need for tailored training interventions

Since interviewer effects are the result of interviewer behaviors, they can be managed by effective training, a finding known for many decades now (Stock and Hochstim 1951; Kish 1962). Billiet and Loosveldt (1988) find that the skills of giving instructions to respondents, probing, and giving feedback are markedly improved with training. Dahlhamer et al. (2010) show that average item response times for some NHIS questionnaire sections show marked improvement after refresher training; training in which considerable time was devoted to reading questions as worded. In reviewing the literature for factors impacting interviewer effects, Schnell and Kreuter (2005) say: "one conclusion seems to apply to most of those studies: interviewer effects can be reduced if the interviewer has received good training, and if the use of a standardized procedure is ensured".

Professionally-run surveys conduct well-structured interviewer training sessions. However, even relatively intensive training of interviewers before a survey does not appear to keep interviewers from performing inadequately during fieldwork (van der Zouwen and Dijkstra 1988), suggesting the need for retraining interviewers during the survey data collection. Two aspects are important here. First, retraining needs to be responsive as early as possible in the interviewing process so that poor interviewer behaviors are corrected early. Second, retraining needs to be tailored; many interviewers may not need a full and expensive retraining but guidance on specific items that they may be struggling with.

These needs are even greater for large-scale surveys given the impact of workload on precision discussed above. Practitioners would like a cost-effective data-driven automatic system to detect interviewing issues that could impact estimate precision; small-scale surveys can sift through data using even semi-manual procedures. However, these kinds of systems do not seem to be common; interviewer effects are rarely evaluated during fieldwork (Biemer and Lyberg 2003, p.168) and interviewers contributing to it rarely identified despite literature calling for it (e.g., West et al. 2013). Two challenges with evaluating interviewer effects during a survey (and therefore interviewers who contribute to it) are that estimates of ρ_{int} are notoriously unstable (Groves 1989, p.368; Biemer and Lyberg 2003, p.168) and the sheer number of items that need to be evaluated would

increase pressure on survey managers who are already dealing with production issues. This is where paradata can potentially play an important role by acting as a proxy for interviewer effects. We use the term 'paradata' in the specific sense that it was first defined by Couper (1998), i.e., keystrokes and time-stamps that are generated in the course of a Computer-Assisted Interview (CAI).

4.1.3 Paradata

Why would we expect paradata to be useful to monitor interviewers for measurement error? Consider two alternate interview scenarios that involve the same respondent but two different interviewers.

Scenario 1

INTERVIEWER1: Thinking back on the past week, on an average, for how many hours a day did you watch TV ?
RESPONDENT: Am not sure.
INTERVIEWER1: Could you guess? For example, about 20 hours?
RESPONDENT: Yes, actually about 20 hours.

Scenario 2

INTERVIEWER2: Thinking back on the past week, on an average, for how many hours a day did you watch TV ?
RESPONDENT: Am not sure.
INTERVIEWER2: Maybe I could repeat the question to help you answer the question. [slows down pace] Thinking back on the past week, on an average, for how many hours a day did you watch TV ?
RESPONDENT: What do you mean by "on average"?
INTERVIEWER2: In the last seven days, there might have been days you could have watched less TV and some days more TV. But if I asked you to give me one number that stands for your daily TV viewing over the week, what would that be? Please take your time to recall your TV viewing last week and answer the question.
RESPONDENT: About two and a half hours.

The two interviews obtained very different responses. Interviewer1 ends up giving a directive probe to the respondent when faced with a potential non-response. On the other hand, interviewer2 undertakes the right steps by first repeating the question to the respondent, clarifying the question in a neutral fashion, and giving encouraging feedback. These actions are likely to have yielded an accurate response. Such interviewer behavior

tends to be consistent as noted by Fowler and Mangione (1990, p.45): "Some interviewers obtain more answers than others on a consistent basis because they consistently probe for more answers, and that affects the data".

Broadly, a respondent goes through the process of comprehending the question (Comprehension), recalling the relevant information from memory (Retrieval), combining various recalled information in order to answer the question (Judgment), and finally communicating the answer (Tourangeau et al. 2000). An interviewer who is speeding through the process could make comprehension difficult, and even if the respondent did comprehend the question, the rushed behavior could give cues that the respondent need not bother responding carefully (Fowler and Mangione 1990, p.71). On the other hand, careful probing would encourage better cognitive processing by the respondent resulting in better data quality. However, this latter effort would also tend to be associated with higher item times. In other words, item times (paradata in general) could be capturing interviewer behaviors that are associated with survey error.

These intuitions go back a long way - e.g., Steinkamp (1964) looks at average interview length and variability in interview length to assess interviewer performance - and are used even in current survey quality control by methods such as focusing on interviews that are completed in less than a threshold time. While such heuristics can be useful, they are inherently subjective and unlikely to be optimal. With technological advancement, the ease of obtaining paradata has increased. They have the advantage of being generated at low marginal cost, are generally not afflicted by missingness, and can be considered relatively error-free (West and Sinibaldi 2013). Moreover, they contain a lot of detail (as seen in Chapter 2). Paradoxically, this feature of paradata has proven to be an obstacle, with analysts often being overwhelmed with the amount of data (Couper et al. 1997; Nicolaas 2011) and finding it challenging to separate signal from noise. Consequently, paradata tend to be highly summarized before they are used, e.g., Biemer and Lyberg (2003, p.429) and Jans et al. (2011). Better use of paradata requires research on its properties but "relatively little attention has been paid to keystroke or item-level paradata" and "the absence of research on the large-scale use of measurement-error-related paradata in interview surveys is unfortunate" (Couper and Kreuter 2013).

We attempt to fill these gaps in this chapter. The overarching goal is to try and harness the power of item-level paradata by linking them explicitly with interviewer effects in order to enable quick, tailored training interventions with interviewers. If successful, it will be a definite advantage over traditional quality control systems that rely on recordings from a small subsample; paradata have no such restrictions, being available on the full sample. Given our intuitions above on how paradata might be predictive of interviewer effects, we preceded this study with separate research (Chapter 2 of this dissertation)

that shows that paradata patterns are associated with indicators of interviewing quality. This helps us proceed with the present research with greater confidence. The surveys we have in mind for our application are repeated cross-section surveys and panel surveys.

We formulate three specific research questions.

1. How effective are paradata in predicting interviewer effects?
2. How do the paradata compare to interviewer demographics, work-related variables, and quality control indicators in predicting interviewer effects?
3. Are there specific paradata variables that analysts should focus on during survey quality control?

The rest of this chapter is organized as follows. Section 4.2 briefly reviews the Panel Study of Income Dynamics (PSID), which is used as the "testbed" for the proposed methods, Section 4.3 discusses in detail the PSID data used in this chapter, Section 4.4 describes the models used to estimate interviewer response variance, Section 4.5 describes the results of the analysis, Section 4.6 discusses the results obtained and gives suggestions for future research, and Section 4.7 considers issues of practical implementation of the proposed methods.

4.2 Study survey

We used data from the 2015 wave of the PSID for our research. The PSID is a nationally representative survey of families and individuals in the U.S., conducted via Computer Assisted Telephone Interviewing (CATI). The survey consists of biennial waves where one respondent per family is administered a 'main interview'; supplemental studies are added to this main interview e.g., the 'transition into adulthood supplement' is asked to individuals when they become 18 years of age. Between March-December 2015, 9048 respondents were interviewed by 96 interviewers with a response rate of 89% (calculated with respect to the previous wave). Interviews were largely conducted by telephone; only 2.8% interviews had to be conducted in face-to-face mode. An interview lasted 80 minutes on average. The detailed questionnaire [1] and codebook [2] are available on the PSID website.

Before the survey commences, PSID interviewers undergo video training [3] on the study terminology, concepts, and individual sections. This is followed by an in-depth in-person

[1] `ftp://ftp.isr.umich.edu/pub/src/psid/questionnaires/q2015.pdf`
[2] `ftp://ftp.isr.umich.edu/pub/src/psid/codebook/fam2015er_codebook.pdf`
[3] `https://psidonline.isr.umich.edu/videos.aspx`

training at Ann Arbor, Michigan, USA. Approximately 60% of the 96 interviewers in PSID 2015 were also interviewers for at least one of the previous two waves.

4.3 Data

4.3.1 Substantive data

The PSID main interview begins by taking consent from the respondent followed by questions about the family composition and member details. These 'coverscreen' questions are followed by substantive questions. On average, a respondent answers about 360 substantive questions across 11 sections as shown in Table 4.1; sections concerning employment (sections BC/DE), expenditures (section F), and health (section H) account for close to 60% of interview duration.

No.	Section	Substantive area	Average # items administered in an IW	Average IW duration (mins)
1	A	Housing, Utilities, Computer Use	36	7
2	BC, DE *(including EHC[1])*	Employment	46	22
3	F	Expenditures	52	11
4	G	Current income; Other family unit member education	48	9
5	R	Off-year income and public assistance	12	2
6	W	Wealth and active savings	22	5
7	P	Pensions	13	3
8	H	Health	96	14
9	J	Marriages and Children	11	1
10	KL	New head and spouse/partner background	14	3
11	M	Philanthropy	9	2
		Average interiew	**358**	**80**

1. EHC: Event History Calendar

Table 4.1: PSID substantive section descriptions.

In 2015, to aid timely research into the aftermath of the 2008 recession, PSID released data on 357 items concerning mortgage distress, housing, food security, wealth, and computer use (belonging to Sections A, F, and W) within a month of fieldwork completion. These early release data (ER data) do not include data from 'split-off' families (split-off families consist of either a person or group of people who moved out from an existing PSID family

since the prior wave's interview to form a new, economically independent family unit living in a separate housing unit). This reduces the total ER data sample size to 8262 families. However, since these data were quickly released, they did not undergo the usual editing, cleaning, and imputing processes which is an advantage for our analytical goals since data 'as collected' would better reflect interviewer effects. We therefore used ER data for these families and the regular public use file data (that have undergone all the data processing) for the remaining respondents. We used only respondents interviewed via CATI for our analysis.

4.3.2 Paradata

We created 13 interviewer-level paradata measures for each item. For count measures, we included the coefficients of variation (CV) along with the mean values; including just the latter would mask variations across interviews that may be important predictors of interviewer effects.

Measures 1.- 4. below are time-based paradata variables. The paradata literature has generally looked at item times as a single measure. But item times are generated by different activities having potentially different correlations with interviewer effects, e.g., asking a question versus recording a response. Given the data we had, we could split an item into 2 parts. The first part counted the number of seconds from the time the interviewer gets into an item on the CATI instrument up to the first keystroke made. This was used as a surrogate for the time taken for the interviewer to Ask the question, Probe and give feedback to the respondent, and Receive a response (abbreviated as 'APR time'). The second part counted the time from the first keystroke up to the exit from the item which we used as a surrogate for Data Entry time ('DE time'). Both APR time and DE time are computed using the first visit to the item when it would actually have been administered to the respondent.

1.-2. APR time (Mean and CV)

3.–4. DE time (Mean and CV)

5.–6. No. of item visits (Mean and CV). From previous research (Chapter 2), we know that multiple visits to an item are an indicator that the interviewer may be facing a problem. A high mean with a low CV for this measure could especially indicate an issue.

7.–10. Keycounts (Mean and CV) and Mouseclicks (Mean and CV) could be important indicators of response editing.

11. Proportion of remarks. This measure and the one below (help access) were included based on research (Chapter 2) that showed that they are correlates of interviewing quality.

12. Proportion of help access

13. Proportion of error messages. During the interviewing process, the CATI software triggers error messages when data which are logically inconsistent or beyond preset numerical ranges are entered.

All measures were top-coded at the 97.5th percentile to prevent very large values from overly influencing our models. For binomial and multinomial response items, we did not use the two keycount measures; any keystroke would largely be due to remarks which would be captured by that measure. For numeric response items, we did not use the CV of mouseclicks but we retained the mean of this measure since some interviewers might indulge in idle clicking when the interview is in progress, a potential indicator of distracted interviewing. Since these measures are item-dependent, their descriptive statistics are presented for a select set of analysis items later in Section 4.5.1.

4.3.3 Interviewer characteristics

We used 3 interviewer demographic variables and 3 variables derived from interviewers' work characteristics.

1. Interviewer sex (88% of interviewers are female).

2. Interviewer age (mean: 53.6 years, standard deviation: 12.1 years)

3. Interviewer education (less than High school, 12% of interviewers; High school/GED, 35% of interviewers; some college, 28% of interviewers, college graduate and above, 25% of interviewers).

4. Interviewer workload, i.e., number of conducted interviews (mean: 114.5 interviews, standard deviation: 41.6 interviews).

5. Mean interviews per day (mean: 1.2 daily interviews, standard deviation: 0.12 daily interviews); even a moderate workload may lead to interviewer fatigue if completed in a short time period. We include partial interviews in this calculation.

6. The coefficient of variation (CV) of the number of daily interviews conducted (mean: 0.53, standard deviation: 0.11); interviewers who are more consistent with the number of daily interviews may be associated with better interviewing quality.

We refer to these 6 interviewer characteristics collectively as Interviewer Demographic and Work-related (IDW) variables.

4.3.4 Interviewing evaluation data

In 2015, PSID recorded two of the first four interviews in every interviewer's workload followed by a further 10% random sample, resulting in 1120 recorded interviews. A 'capture list' dictated which item in the interview was to be recorded; these items were chosen on the basis of substantive importance. For the first three weeks of fieldwork, the capture list inadvertently contained 1157 items belonging to a pretest version. This was corrected and the list pared down to 382 items. We only consider items from the 382 items for our analyses.

Of the 1120 interviews, 594 CATI interviews (53% of all the recorded interviews) were listened to by nine quality control (QC) evaluators. Owing to issues such as bad recordings or missing interviewer characteristics, only 555 interviews were available for analysis. These interviews were conducted by 92 interviewers (96% of the 96 interviewers), with a median of 6 evaluated interviews per interviewer (first quartile: 4 interviews, third quartile: 8 interviews). The recorded items accounted for a median 35% of the total number of administered substantive items (IQR: 31% - 40%) and a median 45% of the substantive interview duration (range: 40% - 50%) within the 555 evaluated interviews.

Apart from their training and extensive experience in behavior coding, many of the QC evaluators have been interviewers themselves which especially equips them to understand interviewer behavior. An evaluator raised a QC flag for an item if she encountered an issue in any of the five interviewing dimensions in Table 4.2. QC flags were classified as 'major' or 'minor' depending on the potential impact on the substantive response.

Table 4.2: The five interviewing evaluation dimensions with sixteen categories.

No.	Interviewing dimension	Categories
1	Question asking	Altered wording; Skipped question; Question delivery; Not verbatim; Other reading error
2	Probing and clarifying	Failure to probe or clarify; Inappropriate, evaluative, or directive probe; Other probing error
3	Data entry	Wrong category; Wrong entry
4	Feedback	Emotive feedback; Other feedback error
5	Other reasons	Unprofessional conduct; Consent error; Household composition; Other error

Coders also explicitly noted the cause of a major flag; four of the sixteen categories accounted for 70% of all major flags: failure to probe or clarify (44%), altered wording (11%), inappropriate, evaluative, or directive probe (9%), and 'other entry error' (6%). The large proportion of flags due to improper probing is expected since interviewers find

it the hardest skill to learn (Fowler and Mangione 1990, p.44); Hicks et al. (2010) find that interviewers probed only in 57 percent of the instances when a probe was needed.

Based on these data, we created the following two interviewer-level 'flag variables' to use in our models:

1. *ItemFlagProportion*, an interviewer-level variable that represents the proportion of evaluated cases for a specific item (across interviews) whch have either a major or minor flag. Since there were only a median 6 evaluated cases per item per interviewer, splitting this variable on the basis of major and minor flags (let alone specific behaviors) was not possible. Figure 4.1 plots the item flag proportions for 27 items in the PSID questionnaire (the choice of these items is described below in Section 4.4.2) where we can see a fair amount of variation in these proportions across items.

2. *OverallFlagProportion*, which is the overall proportion of evaluated cases for an interviewer (across items and interviews) that have either a major or minor flag. The median for this variable is 0.033 with values ranging from 0.008 to 0.099 (IQR: 0.021 - 0.053).

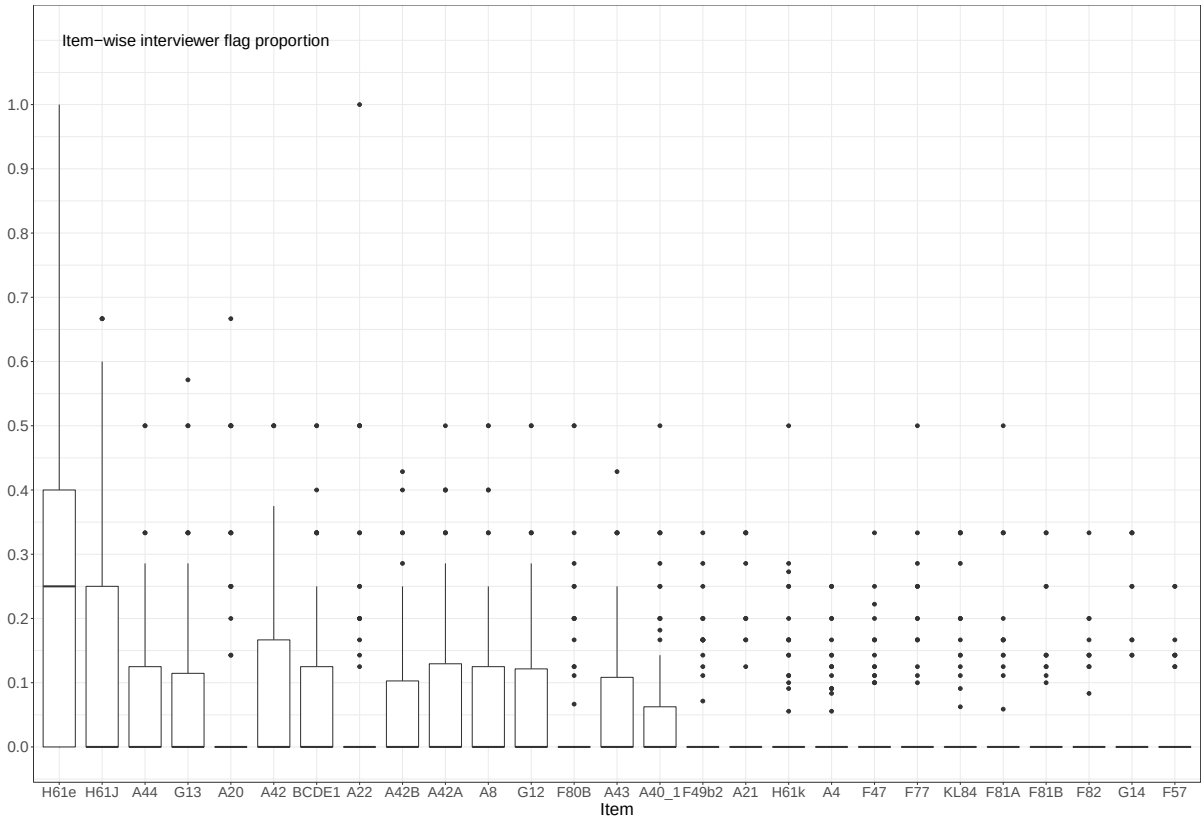

Figure 4.1: Item-wise interviewer flag proportions. Each data point in this plot is the item flag proportion for a specific interviewer. Items on the horizontal axis are sorted in descending order of the flag rate. Items with smaller flag proportions are driven by only a few interviewers.

4.4 Methods

4.4.1 Models for interviewer response variance

Consider the following model with interviewer-varying intercepts fit to y_{ij} responses of a certain continuous item, where the subscript j refers to a respondent who is interviewed by interviewer i.

$$y_{ij} = \beta_0 + u_{0i} + \mathbf{X_{ij}^T}\boldsymbol{\beta_X} + \epsilon_{ij} \tag{4.2}$$

$$u_{0i} \overset{iid}{\sim} N(0, \sigma^2_{iwer})$$

$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2_e)$$

$$u_{0i} \perp \epsilon_{ij}$$

In this model, $\mathbf{X}$ is a vector of respondent covariates via which we seek to approximate an interpenetrated design; we do not have substantive interest in them. We included three household-level variables and three individual-level variables in $\mathbf{X}$: number of adults in the family (1, 2, 3, and 4+ adults) was included to account for possible higher income and expense response values in larger families. Also, many questions in the PSID questionnaire are repeated for every adult in the home, potentially adding to respondent and interviewer burden, thereby increasing response error; number of children at home (0, 1, 2, 3, and 4+ children) was included to account for potentially larger dwelling units and more expenses; reported 2014 household income ($\leq \$25K$, $\$25K - \$50K$, $\$50K - \$75K$, $\$75K - \$100K$, and $> 100K$) was included since the survey is primarily economic in nature and response values for many items would be correlated with the household economic condition; sex was included since past research suggests that this may be associated with recall accuracy (Skowronski and Thompson 1990; Auriat 1993) and measurement error on economic data (Lee and Lee 2012); education (some high school or less, high school graduate, come college, and college graduate and above) was included since this is known to be correlated with cognitive sophistication (Krosnick and Alwin 1987, Krosnick 1991); and finally, age ($\leq$ 25 years, 25-34 years, 35-54 years, 55-64 years, 65-74, and 75 years and above) was included since it is known to account for response effects even after taking into account respondent education (Knauper 1999).

The categorization of continuous control variables helped overcome initial problems during model estimation. It also ensured that outlier values do not overly influence our inferences. We checked if some control variable categories were concentrated only among only a few interviewers but this was not the case. These variables also have low item

missing rates, the highest being that of total income at 1.4%.

Our interest is in $\hat{\sigma}^2_{iwer}$, the estimate of interviewer measurement error variance. We test its statistical significance (at a 0.05 level) using a 50:50 χ^2 approach (Self and Liang 1987). If the interviewer variance component is significant, we we fit the following 'full model' where $\mathbf{P}$ and $\mathbf{Z}$ are the vectors of paradata variables (described in Section 4.3.2) and non-paradata variables (described in Sections 4.3.3 and 4.3.4). The flag proportions that are part of $\mathbf{Z}$ were computed from the QC evaluation data and used as an estimate of the full sample flag proportions.

$$y_{ij} = \beta'_0 + u'_{0i} + \mathbf{X^T_{ij}}\boldsymbol{\beta'}_X + (\mathbf{P_i}^T\boldsymbol{\beta}_P + \mathbf{Z_i}^T\boldsymbol{\beta}_Z) + \epsilon'_{ij} \tag{4.3}$$

$$u'_{0i} \overset{iid}{\sim} N(0, \sigma'^2_{iwer})$$

$$\epsilon'_{ij} \overset{iid}{\sim} N(0, \sigma'^2_e)$$

$$u'_{0i} \perp \epsilon'_{ij}$$

Given our research questions, we wanted to evaluate the performance of $\mathbf{P}$ and $\mathbf{Z}$ in explaining $\hat{\sigma}^2_{iwer}$. We used the proportion of interviewer variance explained, $\hat{p}_{ExplVar} = (\hat{\sigma}^2_{iwer} - \hat{\sigma}'^2_{iwer})/\hat{\sigma}^2_{iwer}$, as our evaluation measure. We compare this measure for the paradata model ('P model', that only adds $\mathbf{P}$ to equation 4.2), the non-paradata model ('NP model', that only adds $\mathbf{Z}$ to equation 4.2), and the full model.

Further, we divide $\mathbf{P}$ into 2 variable blocks: time-based measures and non-time-based measures. The paradata literature has focused on the former with little attention to the latter; separate analyses will provide evidence of the comparative utility of these 2 variable sets. Variables in $\mathbf{Z}$ can be divided into 2 blocks: variables involving interviewer characteristics and work-related variables (Section 4.3.3), and the flag variables (Section 4.3.4). While not all survey organizations undertake the kind of detailed QC that PSID conducts, they will all have the IDW variables which can serve as a minimum benchmark against which to evaluate the paradata. Based on these variable blocks, we fit the following subset models:

1. P-Time, that uses only time-based paradata variables.
2. P-NonTime, that uses only the non time-based paradata variables.
3. NP-IDW, that uses only the IDW variables.
4. NP-Flag, that uses only the flag variables.

The P and NP models are formed by adding terms from their respective subset models.

Model for a binary response item.
When y_{ij} is a binary variable with $y_{ij} \sim BER(p_{ij})$, we fit logit models where the predictor part is structurally similar to the linear models above (the same coefficients have been retained for simplicity). The base and full models are as follows.

$$log(\frac{p_{ij}}{1-p_{ij}}) = \beta_0 + u_{0i} + \mathbf{X_{ij}^T}\boldsymbol{\beta_X} \tag{4.4}$$

$$u_{0i} \overset{iid}{\sim} N(0, \sigma^2_{iwer})$$

$$log(\frac{p_{ij}}{1-p_{ij}}) = \beta'_0 + u'_{0i} + \mathbf{X_{ij}^T}\boldsymbol{\beta'}_X + (\mathbf{P_i}^T\boldsymbol{\beta}_P + \mathbf{Z_i}^T\boldsymbol{\beta}_Z) \tag{4.5}$$

$$u'_{0i} \overset{iid}{\sim} N(0, \sigma'^2_{iwer})$$

For multinomial items, we fit separate logistic models to each category that constituted at least 5% of the responses. Doing so, rather than fitting a single multinomial model, gives us the flexibility to fit models with potentially different forms for each category. While not technically accurate, we refer to these individual categories as items too for descriptive convenience.

4.4.2 Choosing items for analysis

Since our analysis also involves the flag variables, we first looked at all substantive items with a minimum of 200 interviewing evaluations (approximately a minimum of 2 evaluations each for the 92 interviewers) and at least 10 flags (so we have some potential effects to measure). We then dropped the following items: 'Event History Calendar' (EHC), which is really a battery of questions on employment and residence but which the interviewing evaluation process treats as a single 'item'; two open-ended items on the nature of work and industry (BC20 and BC21); a reading evaluation of the introduction to section M of the questionnaire; a question on which specific member is covered by health insurance (H61D3); and an item on which family member's employer provides insurance (H61f). Finally, we removed binary item 'G17f', due to a very low (only 0.2%) substantive response proportion. This yielded 27 items which are a mix of response types: numeric (14 items, with 13 of them involving monetary values), binary (7 items), and multinomial (6 items). On average, an item in this set was administered to 7276 respondents (standard deviation of 1554 respondents) by 89 interviewers (standard deviation of 4 interviewers). Since categories of multinomial models were modeled separately we had a total of 45 resultant 'items'.

We wanted to select items (we drop the quotes around 'items' unless necessary) with different magnitudes of $\hat{p}^{NP}_{ExplVar}$ so that we can benchmark the performance of the para-data variables at these different levels. Therefore, we first fit the model in equation 4.2 or equation 4.4 to each of the 45 items followed by fitting the NP model to those items with a significant $\hat{\sigma}^2_{iwer}$. We then selected 11 items such that they spanned the range of $\hat{p}^{NP}_{ExplVar}$ and included all response types. We avoided selecting more than one category of the same multinomial variable so as to get more item heterogeneity. The results of this analysis are shown later in Section 4.5.1.

4.4.3 Model fitting and analysis details

Model variable selection

We decided to use an automatic variable selection method since we wanted to discover terms, their higher powers, and interactions that we might miss with a manual variable selection process. Also, we envisage our research being used in practice for many items (e.g., hundreds of items) in a questionnaire in which case fitting models to each item 'by hand' would be impractical. We chose the Adaptive Least Absolute Shrinkage and Selection Operator (ALASSO, Zou 2006) to undertake variable selection. By using the L1 penalty and selecting variables by cross-validation, ALASSO gives us a subset of variables that have predictive importance as well as reduce the chances of overfitting. The procedure also has the 'oracle property' that gives us consistent variable selection and coefficient estimates. We used the *polywog* package [Kenkel and Signorino 2018] in the R [Team 2013] software to undertake ALASSO. Our initial predictor input included all square transforms and two-way interactions. Based on the results in Chapter 3 we also included cubic transformations for the flag variables. We forced back main effects in situations where the algorithm selected interactions or higher powers but did not select the main effects themselves. The ALASSO penalty factor was chosen on the basis of a 10-fold cross-validation(CV); folds were created by grouping interviewers into 10 groups so that each group had approximately the same number of respondents. The control variables ($\mathbf{X}$) were left unpenalized.

The final selected inputs based on ALASSO (selected at a 5% significance level) were then used to fit the multilevel model in equation 4.3 or equation 4.5 . All multilevel models were fit with the *lme4* package (Bates et al. 2015) using the Laplace approximation in the R (Team 2013) software. For linear multilevel models, we used Restricted Maximum Likelihood (REML) that was fit using the Bobyqa optimizer. Logistic models in *lme4* use ML estimation. Here, we used the 'NLopt' implementation of the BOBQYA optimizer (Johnson) via its R interface 'nloptr' (Ypma et al. 2014); some testing showed that variance estimates were almost exactly the same as when the default BOBYQA optimizer

was used but with more than a 50% reduction in runtime. The reduction in runtime was important since our model validation methods involve bootstrapping (explained below).

All numeric input variables were centered and scaled to facilitate convergence of the model estimation process. We do not screen for 'outliers' when running our models since these are actually potentially valuable in being responsible for interviewer effects. We do not include survey weights since we are interested in uncovering interviewer effect structures conditional on the current PSID design, not averaged over the design to the population. Since all our variables of interest are at the interviewer-level, coefficient rescaling that is recommended to make coefficients comparable across models (Snijders and Bosker 1999, p. 228-229; Hox 2010, p.134; Austin and Merlo 2017) is not necessary to be undertaken (Hox 2010, p.138). Approximately 3% of all interviews were undertaken by multiple interviewers. For these interviews, we used paradata to match an item to the interviewer who actually asked the question.

Model diagnostics

We undertake simulation-based diagnostics as described by Hartig (2018) for all our models. The key idea is that data simulated from the fitted model should mimic the observed data if the fitted model was correctly specified (Gelman and Hill 2006, p. 158-159). To do this, a thousand datasets are simulated from the model, conditioning on all random effects. Then, for each observation a quantile residual (Dunn and Smyth 1996) - defined as the proportion of simulated values larger than the observed value - is computed and two plots are constructed as described below.

- If there are no model fit issues, we would expect the quantile residuals across observations to be uniformly distributed. We draw a quantile-quantile plot to evaluate this; more formally, a Kolmogorov-Smirnov test is conducted to detect deviation from uniformity.

- The quantile residuals are plotted against the mean simulated value for each observation (similar to the diagnostic plot of residuals versus fitted values constructed for a linear model). The mean simulated values are rank transformed and scaled to make it easier to spot issues. To make the analysis more concrete and help protect against missing patterns visually (especially when there are a lot of observations), a quantile regression is conducted between the 25th percentile, median, and 75th percentile of the mean simulated values and the quantile residuals; the quantile regression lines should ideally match horizontal lines at these percentiles that would indicate no association between the residuals and the mean simulated responses. The quantile regression is conducted using quantile regression neural network models via the QRNN package (Cannon 2011) in R, so as to be able to spot potential non-linearities in the patterns.

R code for undertaking the simulation-based analyses was adapted from the source code of the DHARMa package (Hartig 2018) and is included in Appendix 4.B.

Predictive evaluation

Though cross-validation is used to select variables, evaluation of our models' predictive ability will still be optimistic since we are using the same data for selection and evaluation. To get a more realistic measure of predictive ability, we conducted the following steps.

1. We generated 200 bootstrapped datasets (the number of resamples is based on recommendations in Harrell et al. 1996) by resampling interviewers.

2. All our models are fit to each resample using the variable selection procedure described. We use the subscript 'Bootstrap' in place of $\hat{p}_{ExplVar}$ to denote such estimates.

3. The variables selected when fitting each of the bootstrap models are used to then fit corresponding models to the original data. We use the subscript 'Predicted' in place of $\hat{p}_{ExplVar}$ to denote such estimates.

4. Since, on average, approximately 37% of the data is not present in each bootstrapped sample (Efron 1983; Efron and Tibshirani 1997), the average $\hat{p}_{Predicted}$ will give us a more realistic measure of model performance.

We use the subscript 'Apparent' in place of 'ExplVar' in $\hat{p}_{ExplVar}$ when we refer to estimates computed when variable selection as well as model fitting are conducted on the original data. We use superscripts to denote the model being fit, e.g., $\hat{p}^{P-Time}_{Apparent}$.

4.5 Results

4.5.1 Items selected and descriptive analysis

Of the 45 items analyzed, we found that the interviewer variance component was significant for all but 6 items (F81B, F80B, G13, F82, KL84, A4_2; 'A4_2' indicates category code 2 of multinomial item A4). Of the remaining 39 items, the adaptive lasso algorithm failed to converge for items H61e_12 and BCDE1_6. Table 4.3 shows $\hat{p}^{NP}_{ExplVar}$ for the remaining 37 items along with their descriptions. For a large number of items (22 items), the non-paradata variables failed to explain any of the estimated interviewer variance. While these results are not directly comparable to those obtained in Chapter 3 since the model building methods are different, they are directionally the same.

The highlighted rows in Table 4.3 correspond to the 11 items that we chose for further

analysis. We conducted a descriptive analysis before we fit the P models to these items. For each paradata measure across items, the mean magnitude and ± 1 standard deviations across interviewers are plotted in Figures 4.2 (time and item visit measures) and 4.3 (all other measures). Some measures are strongly prevalent for some items, e.g., proportion of help for item A8 (number of rooms), or exhibit more variation for specific items, e.g., proportion of remarks for item A44 (Telephone/TV/Internet expenses); our formal models will check if these between-interviewer paradata differences are predictive of between-interviewer response variance. Binomial and multinomial response items generally have a lower DE time than the numeric response items since they potentially only involve a mouseclick. However the presence of a non-zero mean DE time for these items indicate that one or more of the following activities are taking place even after the initial response is entered: response editing, accessing help, entering remarks, interactions with the respondent, etc. Descriptive statistics for the 11 items based on the substantive responses are given in Appendix 4.A.1.

Table 4.3: The original item pool along with the 11 selected items. Data are sorted by $\hat{p}^{NP}_{Apparent}$. The highlighted rows correspond to the 11 items that were finally chosen for the analysis. The (B), (M), and (N) in the third column refer to binomial, multinomial, and numeric response for that item; the underscore appended to multinomial items is the category number that is being modeled.

No.	Item	Description	$\hat{\sigma}^2_{iwer}$	$\hat{p}^{NP}_{Apparent}$
1	A4_4	(M) Dwelling - Mobile home/Trailer	0.33	0.80
2	A40_1_3	(M) Home heating - Oil	0.38	0.67
3	A22	(N) Yearly homeowner's insurance premium	6173.78	0.59
4	A40_1_10	(M) Home heating - Bottled gas or propane	0.36	0.47
5	G12	(B) Earn wages/salaries apart from uninc. business?	0.04	0.43
6	A4_6	(M) Dwelling - Row or town house	0.38	0.43
7	BCDE1_3	(M) Employment status - looking for work, unemployed	0.05	0.42
8	F49b2_4	(M) Vehicle type - Alternate fuel	0.55	0.41
9	A21	(N) Yearly property taxes	$3x10^5$	0.37
10	A43	(N) Monthly water and sewer amount	56.97	0.33
11	G14	(B) Any income from bonus, tips, commissions?	0.12	0.28
12	A42_1	(M) Utility bill type - One bill and can separate amounts	0.18	0.23
13	A4_3	(M) Dwelling - Apart. in multi-unit building	0.06	0.22
14	H61J	(N) Monthly health ins. premium	328.63	0.17
15	A20	(N) Present home value	$1.1x10^9$	0.17
16	F49b2_5	(M) Vehicle type - none of the options	0.22	0.16
17	BCDE1_1	(M) Employment status - working now	0.02	0.14
18	A42_2	(M) Utility bill type - One bill and can separate amounts	0.14	0.12
19	H61e_5	(M) Medicaid insurance	0.06	0.12
20	H61k	(B) Was anyone without insurance?	0.04	0.11
21	A4_1	(M) Dwelling - One-family house	0.05	0.09
22	F77	(N) Amount spent on car insurance	198.91	0.09
23	A44	(N) Telephone/TV/Internet expenses	275.96	0.08
24	A42B	(N) Amount spent on electricity	214.53	0.06
25	F47	(B) Own/lease car for personal use?	0.10	0.02
26	H61e_3	(M) Medicare insurance	0.10	0.02
27	A40_1_1	(M) Home heating - Gas	0.09	0.00
28	A40_1_2	(M) Home heating - electricity	0.12	0.00
29	A40_1_97	(M) Home heating - Other specify	0.59	0.00
30	A42_5	(M) Do not get one utility bill	0.13	0.00
31	A42A	(N) Amount on gas or other fuel for home	682.87	0.00
32	A8	(N) No. of rooms	0.10	0.00
33	BCDE1_5	(M) Employment status - disabled	0.15	0.00
34	F49b2_1	(M) Vehicle type - Hybrid	0.29	0.00
35	F57	(B) Vehicle used for business purposes?	0.10	0.00
36	F81A	(N) Amount on transportation last month	10.16	0.00
37	H61e_1	(M) Employer provided insurance	0.18	0.00

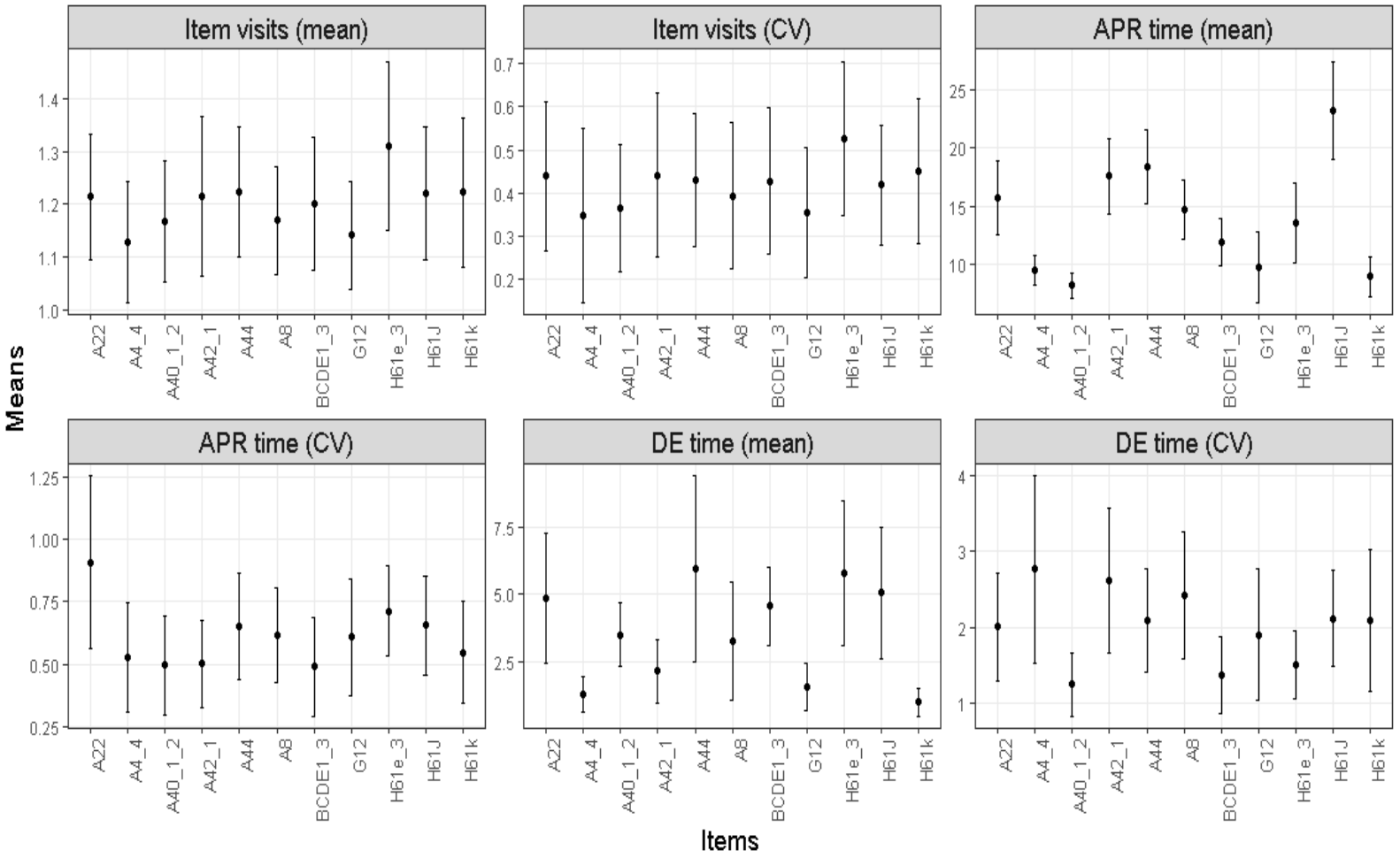

Figure 4.2: Means and standard deviations for the time and item visit paradata measures. The 11 items (horizontal axis in each panel) are as follows where B, N, and M denote binomial, numeric, and multinomial response types: A22 (Yearly homeowner's ins. premium, N), A4_4 (Dwelling - Mobile home/Trailer, M), A40_1_2 (Home heating - electricity, M), A42_1 (One utility bill and can separate amounts, M), A44 (Telephone/TV/Internet expenses, N), A8 (No. of rooms, N), BCDE1_3 (Employment status - looking for work, unemployed, M), G12 (Earn wages/salaries apart from uninc. business?, B), H61e_3 (Medicare insurance, M), H61J (Health insurance premium, N), H61k (Was anyone without insurance?, B)

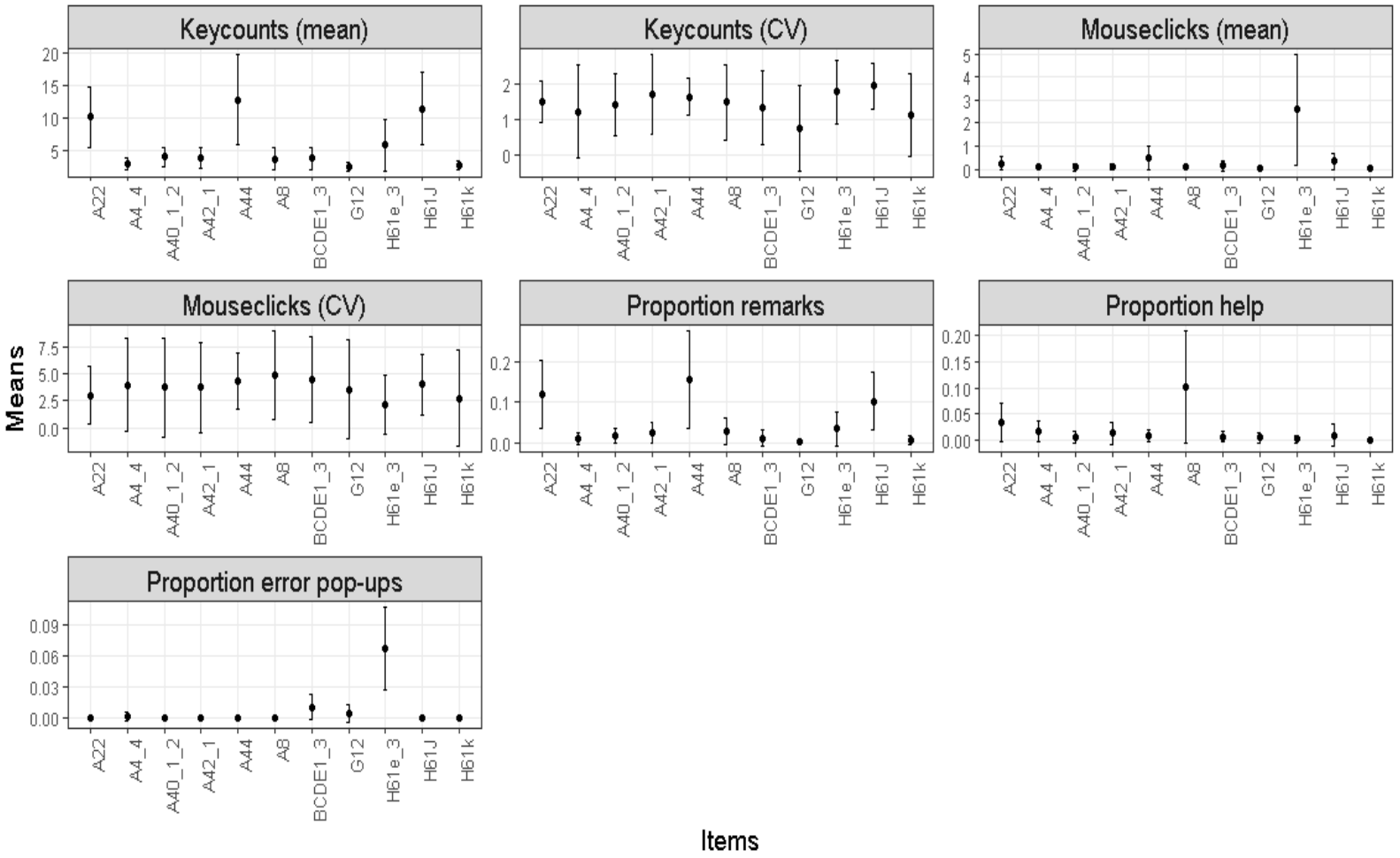

Figure 4.3: Means and standard deviations for the non-time and non-item visit paradata measures. The 11 items (horizontal axis in each panel) are as follows where B, N, and M denote binomial, numeric, and multinomial response types: A22 (Yearly homeowner's ins. premium, N), A4_4 (Dwelling - Mobile home/Trailer, M), A40_1_2 (Home heating - electricity, M), A42_1 (One utility bill and can separate amounts, M), A44 (Telephone/TV/Internet expenses, N), A8 (No. of rooms, N), BCDE1_3 (Employment status - looking for work, unemployed, M), G12 (Earn wages/salaries apart from uninc. business?, B), H61e_3 (Medicare insurance, M), H61J (Health insurance premium, N), H61k (Was anyone without insurance?, B)

4.5.2 Modeling results

We describe our results in top-down fashion by starting with comparing performances of the P and NP models. We then proceed to explore the components of the P model, namely, the P-time and P-NonTime models, including conducting coefficient-level analysis for these subset models. We finally come back to study possible complementarities between the P and NP models.

Comparison of NP and P models

We compare the apparent performances of the P and NP models in Figure 4.4. The NP model estimates in this figure are the same as shown in Table 4.3. For 7 of the 11 items, the P model outperforms the NP model and quite substantially so; the difference in $\hat{p}_{Apparent}$ between the P and NP models for these items is 0.39 with the P model explaining an average 52% of $\hat{\sigma}_{iwer}$.

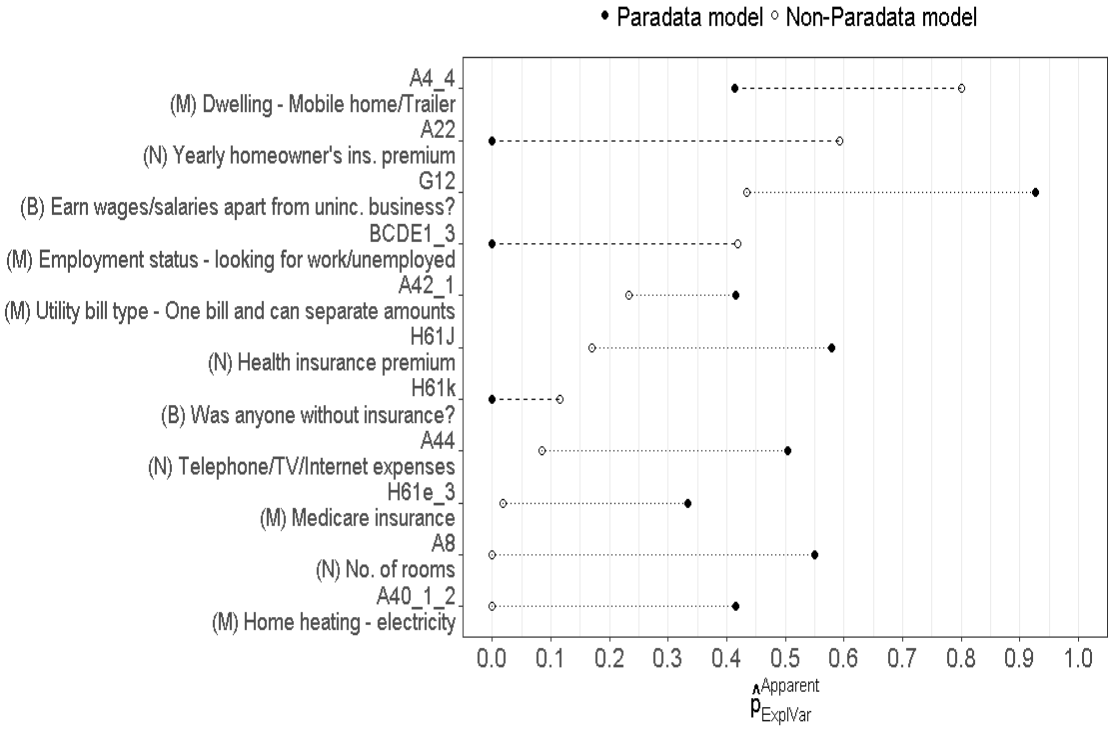

Figure 4.4: Comparison of $\hat{p}_{ExplVar}$ for the P and NP models. The P models are shown by filled circles and the NP models are shown by empty circles. Items are sorted in descending order of $\hat{p}^{NP}_{Apparent}$. The (B), (M), and (N) in the item labels on the vertical axis indicate binomial, multinomial, and numeric response types respectively. The connector lines between the NP and P estimates for an item are only visual guides to judge differences in estimates; the dashed lines denote the 4 cases when $\hat{p}^{NP}_{Apparent}$ is greater than $\hat{p}^{P}_{Apparent}$.

While these are encouraging results, as explained earlier they are optimistic due to which we conducted the bootstrapped-based prediction analysis. We did not include bootstrap

resamples for which problems in estimation or computing were encountered for any model. Yet, 9 of the 11 items had at least 195 valid resamples; items H61K (was anyone without insurance) and G12 (whether earned any salary apart from unincorporated business) had 188 and 180 valid resamples respectively. Since all interviewers do not conduct the same number of interviews, resample sizes were not constant. However, the coefficients of variation were small (range: 3.7% - 4.3% across items).

Figure 4.5 plots the $\hat{p}_{Predicted}$ for the P and NP models along with the interquartile (IQR) ranges.

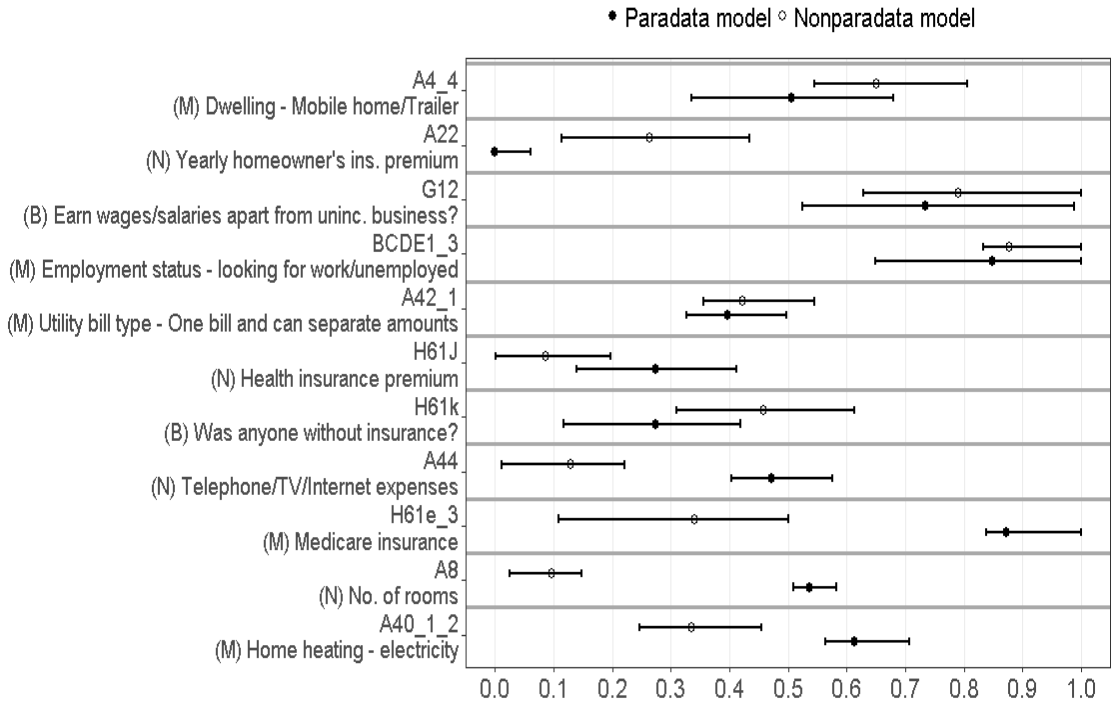

Figure 4.5: Comparison of $\hat{p}_{Predicted}$ for the P and NP models. The P models are shown by filled circles and the NP models are shown by empty circles. Items are sorted in the same order as Figure 4.4. The (B), (M), and (N) in the item labels in the vertical axis indicate binomial, multinomial, and numeric response types respectively. The horizontal lines with each point are the prediction IQRs.

Our general conclusions remain the same. For 5 items - H61J (health insurance premium), A44(telephone/Internet expenses), H61e_3 (whether Medicare insurance availed), A8 (number of rooms), and A40_1_2 (home heating - electricity) - the P model continues to do better than the NP model with an average $\hat{p}_{Predicted}$ difference of 0.36. We also see the prediction IQRs generally well separated between the P and NP models for these items. For 3 items - A4_4 (Dwelling - Mobile home/Trailer), A22 (Yearly homeowner insurance premium), and H61K (anyone without insurance) - the NP model continues to do better than the P model with an average $\hat{p}_{Predicted}$ difference of 0.20. For the remaining 3 items - G12 (Earn wages/salaries apart from unincorporated business), BCDE1_3 (looking for work, unemployed), and A42_1 (One utility bill and can separate amounts)

- the $\hat{p}_{Predicted}$ for the P and NP models are very close to each other.

A counter-intuitive result is that $\hat{p}_{Predicted}$ is greater than $\hat{p}_{Apparent}$ for some items. A stark example is that of item BCDE1_3 (looking for work, unemployed) where $\hat{p}^P_{Apparent}$ is 0 while $\hat{p}^P_{Predicted}$ is 0.85. To get a better sense of the optimism inherent in our estimates, Figure 4.6 plots the optimism indices for the P and NP models which are computed as $\hat{p}_{Bootstrap} - \hat{p}_{Predicted}$. We find that the predictions are subject to a fair amount of optimism; the average optimism index is 0.3, which is the same for the P and NP models (8 of the 11 items are close to the 45 degree diagonal line). However, we see a wide range of values; leaving aside items H61J (Health insurance premium) and A22 (yearly homeowner's premium) that have relatively extreme values, the optimism indices range from 0.1 - 0.5 for the P models and 0.09 - 0.7 for the NP models. A surprising result is that of item H61J where $\hat{p}_{Predicted}$ is better than $\hat{p}_{Bootstrap}$. For item BCDE1_3 (looking for work, unemployed) that we referred to above, the optimism index for the P model is just 0.14, implying that the large difference in $\hat{p}^P_{Predicted}$ and $\hat{p}^P_{Apparent}$ that we saw can be trusted and that we need not always be pessimistic about model fit in predicting interviewer effects.

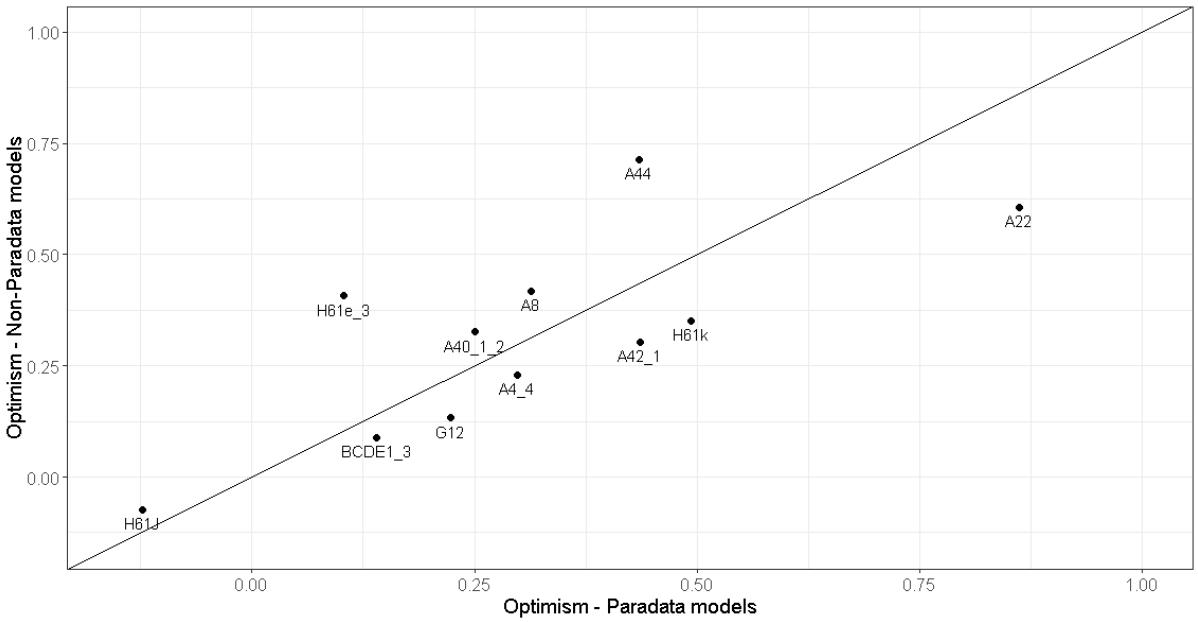

Figure 4.6: Optimism indices for the P and NP models. The indices for the P models are on the horizontal axis and those of the NP models are on the vertical axis. The diagonal line is the 45 degree line.

For the rest of this chapter we will only use $\hat{p}_{Predicted}$ for our inferences since they are conceptually more realistic than $\hat{p}_{Apparent}$. Before we proceed to study the components of the P model, we check if the differences between the P and NP models would have been higher had we used only the IDW variables (and not the flag variables) in the NP models. Figure 4.7 shows that the flag variables add only little over the IDW variables (incremental $\hat{p}_{Predicted}$ of 0.05, on average). For these 11 items, the direction of results is

the same that we obtained in research question 1 of Chapter 3.

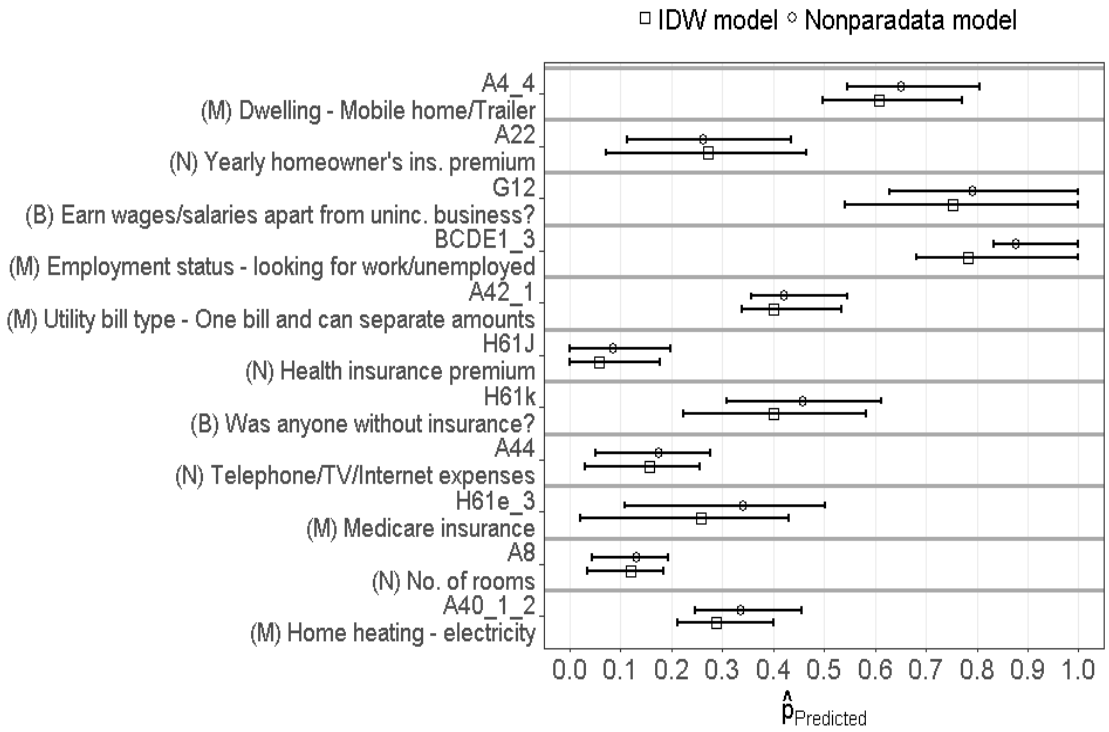

Figure 4.7: Comparison of $\hat{p}_{Predicted}$ for the NP and IDW models. The NP models are shown by circles and IDW models by squares. Items are sorted in the same order as the earlier figures, i.e., in descending order of $\hat{p}^{NP}_{Apparent}$. The (B), (M), and (N) in the item labels on the vertical axis indicate binomial, multinomial, and numeric response types respectively. The horizontal lines with each point are the prediction IQRs.

Comparison of P-Time and P-NonTime models

Figure 4.8 plots $\hat{p}_{Predicted}$ for the P-Time, P-NonTime, and P models. We first focus on the differences between $\hat{p}^{P-Time}_{Predicted}$ and $\hat{p}^{P-NonTime}_{Predicted}$. There is only 1 item - item A8 (number of rooms) - for which the P-Time model does better than the P-NonTime model (mean difference of 0.44). The P-Time IQRs are narrow indicating that the time measures are reliable predictors for this item. For the remaining 7 items (ignoring the 3 items in the top row of the figure for which the differences between the models are too small to consider), the P-NonTime model outperforms the P-Time model with a mean $\hat{p}_{Predicted}$ difference of 0.27 (range: 0.13 - 0.56). However, the P-NonTime IQRs tend to be wider than those of the P-Time models; more non-time variables add to predictive power but this comes at the cost of predictive uncertainty.

We now check if the time and non-time variables explain the same source of variance by comparing the P-Time and P-NonTime estimates to the paradata model estimates in Figure 4.8. Leaving aside item A22 (yearly home insurance premium) for which $\hat{p}^{P}_{Predicted}$

is negligible, we find 3 items - A44 (telephone/TV/Internet expenses), A8 (number of rooms), and H61e_3 (whether Medicare insurance) - where the time and non-time variables subsume the other, as the case may be, in explaining $\hat{\sigma}^2_{iwer}$. For the other 7 items, we can see that the time and non-time variables complement each other in explaining $\hat{\sigma}_{iwer}$; the incremental $\hat{p}_{Predicted}$ over the larger of the 2 estimates is 0.14 (range : 0.10 - 0.18).

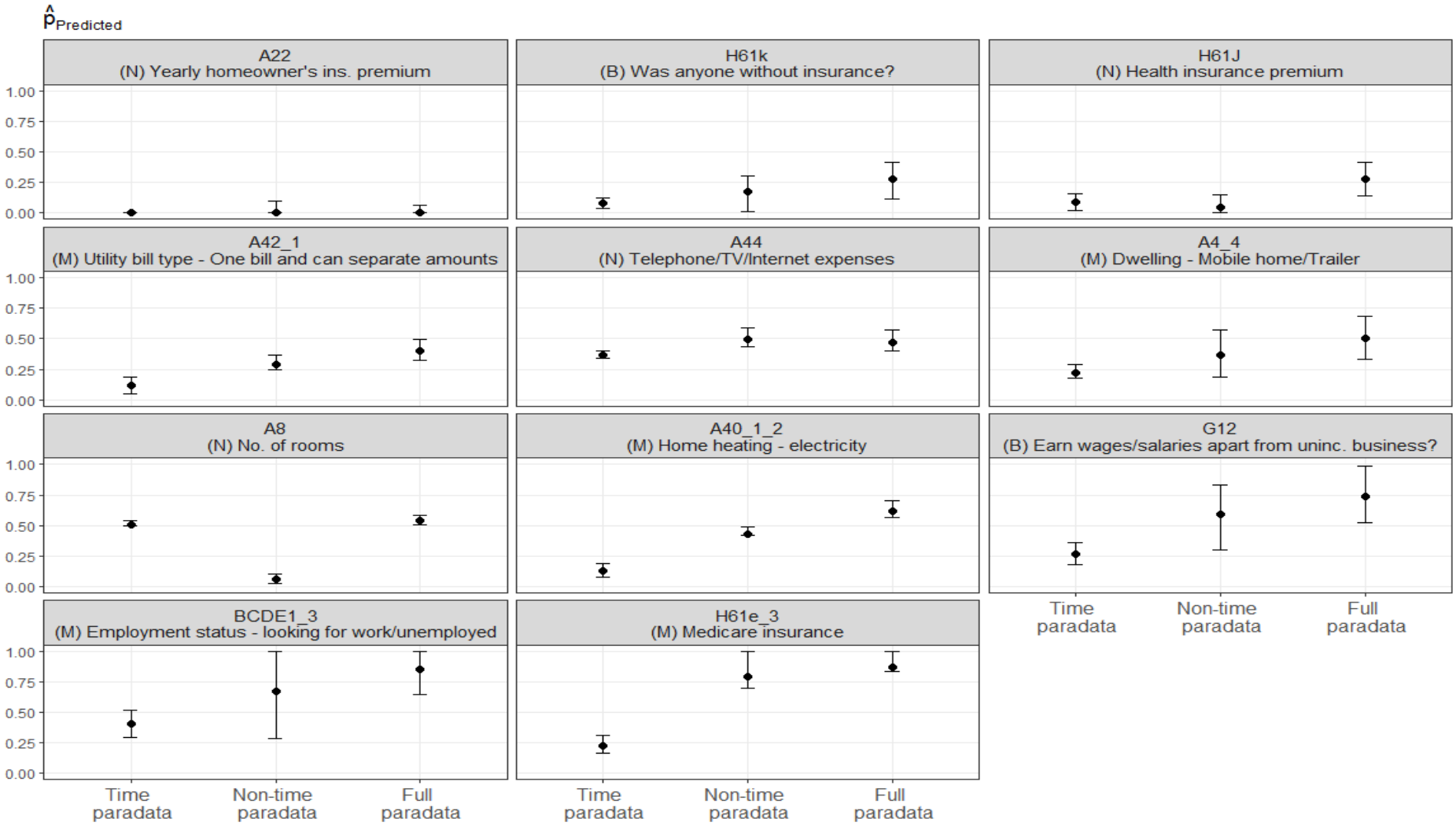

Figure 4.8: Comparison of P-Time, P-NonTime, and P model performances. Each panel contains estimates of $\hat{p}_{Predicted}$ for the three models along with the IQRs (lines running through the points). Panels are arranged in ascending order of $\hat{p}^{P}_{Predicted}$. The (B), (M), and (N) in the panel labels indicate binomial, multinomial, and numeric response types respectively for the associated items.

Coefficient analysis: P-Time model

What specific paradata measures work best in explaining $\hat{\sigma}^2_{iwer}$ for an item? We gauge this by the proportion of bootstrap models in which a measure occurs, and the mean magnitude and standard deviation of the measure's coefficient in the prediction models. We look at main effects that occur in at least 80% of the bootstrap models. Since we forced main effects back when needed, they will tend to have a higher occurrence; we set the occurrence threshold for quadratic terms and interactions at a lower 60% for their estimates to be displayed.

We take up the P-Time model and focus on the 4 items for which $\hat{p}^{P-Time}_{Predicted}$ was at least 0.25. The results for these are shown in Table 4.4, with the columns arranged in descending order of $\hat{p}^{P-Time}_{Predicted}$. Of the 4 time variables, mean APR time is the most important variable in terms of frequency of occurrence. The CV-based measures show that, after controlling for respondent characteristics, between-interviewer differences in how the measures vary within workloads (and not just differences in means) are predictive of $\hat{\sigma}^2_{iwer}$. The high occurrence of mean DE time for A44 (telephone/TV/Internet expense), compared with the other 3 items, aligns with the descriptive analysis in Figure 4.2 which showed that the strongest between-interviewer variation for mean DE time was for this item. This again reinforces the fact that differences in paradata among interviewers are explaining interviewer effects.

Looking at the coefficients, we see that almost all of them have small standard deviations compared to the means which is a good indicator of predictive stability. All displayed coefficients were also consistent in their sign (not shown in the table); across items, the minimum proportion of resample models with the same sign for any of the time measures was 0.97.

The presence of interactions and higher powers makes it difficult to judge the magnitude and direction of a measure's impact on the response. We therefore choose two measures - mean APR time and mean DE time - and compute predictions of their impact on substantive responses. These mean-based measures rather than their CV counterparts were chosen since they are easier to interpret. The predictions were done at 2 levels of both paradata measures - 2 standard deviations above ('High') or below ('Low') the mean - giving us 4 scenarios in all. We assume all other measures are held constant. The results are displayed in Figure 4.9. Results for items BCDE1_3 and G12 are in terms of log-odds while for A8 and A44, they are in terms of the numeric values (number of rooms and monthly dollar expense on telephone/TV/Internet, respectively)

Looking at the first panel, it is clear that the distinguishing measure for item A8 (number of rooms) is mean APR time. On average, when we have a speeding interviewer (Low

Table 4.4: Occurence proportions and coefficient estimates of the time paradata variables. These are proportions of times a time paradata measure occurs in the bootstrap models along with the mean coefficient estimates for the 4 items for which $\hat{p}^{P-Time}_{Predicted}$ is greater than 25%. Columns are arranged in descending order of $\hat{p}^{P-Time}_{Predicted}$. Only main effects with an occurrence proportion greater than 0.8 are shown; the minimum threshold for quadratic terms and interactions is 0.6. The top 2 main effects for each item are in bold. The B, M, and N in the item labels indicate binomial, multinomial, and numeric response types for those items. The $\hat{\beta}$ correspond to scaled measures used as inputs in the models. The models for items A8 and A44 are linear multilevel models predicting the no. of rooms and monthly telephone/TV/Internet expenses, respectively. The models for items BCDE1_3 and G12 are logistic multilevel models predicting the log-odds of looking for work and being unemployed, and whether wages/salaries apart from unincorporated business are being earned, respectively.

	A8 (No. of rooms, N)		BCDE1_3 (Empl. status - looking for work, unemployed)		A44 (Telephone/TV/ Internet expenses)		G12 (Earn wages/salaries apart from uninc. business?)	
$\hat{p}^{P-Time}_{Predicted}$	51%		41%		36%		26%	
	Occurrence (proportion)	Mean $\hat{\beta}$ (sd)	Occurrence (proportion)	Mean $\hat{\beta}$ (sd)	Occurrence (proportion)	Mean $\hat{\beta}$ (sd)	Occurrence (proportion)	Mean $\hat{\beta}$ (sd)
Main effects								
APR time (mean)	**1.00**	0.2 (0.02)	**0.81**	0.07 (0.02)	0.87	4.5 (1)	**0.92**	0.08 (0.03)
DE time (mean)	0.84	-0.04 (0.02)	0.74	0.05 (0.01)	**0.97**	-9.4 (1.6)		
APR time (CV)	0.86	-0.05 (0.04)	**0.80**	0.06 (0.02)			**0.83**	-0.05 (0.02)
DE time (CV)	**0.98**	-0.11 (0.02)	0.69	0.05 (0.02)	**0.90**	9.1 (4.4)		
Quadratic terms								
APR time (mean)	0.68	-0.09 (0.01)						
DE time (mean)					0.63	5 (1.7)		
APR time (CV)	0.84	-0.1 (0.01)						
DE time (CV)							0.67	0.07 (0.02)
Interactions								
APR time (mean) : APR time (CV)							0.72	-0.14 (0.01)
APR time (mean) : DE time (CV)	0.92	0.17 (0.01)			0.61	7.6 (1.1)	0.71	-0.09 (0.02)
DE time (mean) : DE time (CV)					0.63	10.7 (3.1)		

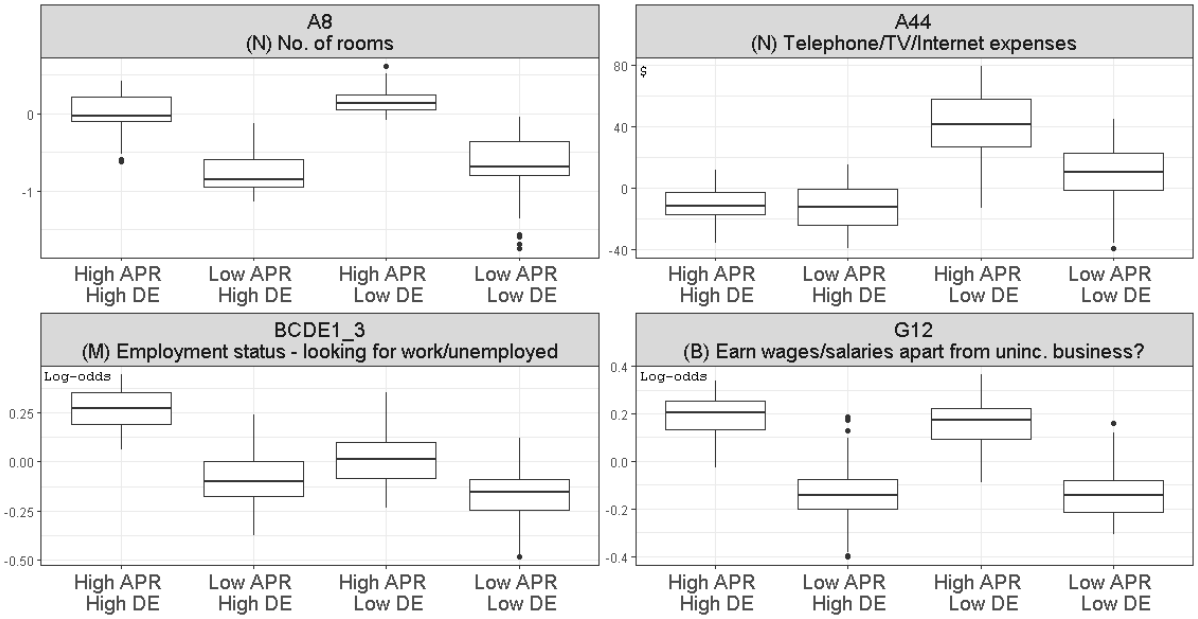

Figure 4.9: Response predictions based on the time paradata measures, e.g., 'High APR' refers to high APR mean time. The boxplots show the variation in changes to the response across the approximately 200 prediction models for the combination of levels shown in the horizontal axis.

mean APR time), irrespective of whether the data entry time is high or low, we can predict an undercount in the number of rooms. But this impact is not symmetrical, i.e., interviewers associated with high mean APRs, on average, are associated with values close to the average; spending a lot of time administering the item may be inefficient.

On the other hand, for item A44 (monthly TV/telephone/Internet expense), mean DE time is the differentiating measure. Low mean DE times are associated with higher dollar expenses perhaps reflecting confidence in the responses. Further, on average, a low DE time coupled with high APR time is associated with a substantial median $40 increase in reported monthly TV/telephone/Internet expenses. But the 2 situations involving high DE time are associated with lower response values. Perhaps, the high DE times are a reflection of a lack of clarity at the interviewer's end which may find expression in activities such as back-forth with the respondent even after the response is recorded, response editing, making a remark etc. These scenarios indicate that paradata are capturing item-specific nuances of interviewer behaviors that lead to interviewer effects, even when the items are of the same response type.

In the case of BCDE1_3, a speeding interviewer (low mean APR time), on average, is associated with a lower odds of an 'unemployed, looking for work' response. This is possibly due to the sensitive nature of the question so that speeding is helping the respondent gloss over possible unemployment by, say, responding with response option 2 which is "Only temporarily laid off, sick or maternity leave". If the respondent does says they s/he is unemployed and looking for work, the special instruction to the interviewer for

this question is to verify if they are looking for work. If 'yes', then response option 3 is indeed the correct answer failing which the interviewer is asked to record the response as option 6 ('keeping house'). Interviewers associated with a high mean APR time (more time probing) and low DE time (confident answers) do not have an impact on the substantive response (median predicted response close to zero). However, a high mean APR accompanied by a high mean DE time results, on average, a markedly higher 32% higher odds of choosing this response option. It is difficult to say whether this is for the better or worse. If the 'extra' DE time is used to probe more after the first response is recorded (perhaps the interviewer gets a hint of discomfort from the respondent), then a more accurate response will perhaps be obtained. On the other hand, using this time to access help, enter remarks etc. maybe a sign of lack of clarity.

In the case of G12 (whether earned wages/salaries apart from unincorporated business), the interviewer is asked to read out the list of employers (from the event history calendar) worked for by the respondent and if necessary review her/his employment history. An interviewer going through this process diligently would give the respondent a good chance of remembering any source of wages/salaries previously forgotten, thus leading to a positive response. This is reflected in the results for G12 in Figure 4.9 where an interviewer associated with a high APR time, on average, is predicted to have an approximately 20% higher odds of getting a 'yes' to this question. Mean DE time does not have an impact for this item (Table 4.4 also showed that this measure was relatively infrequent); Figure 4.3 shows that the mean DE time is approximately 1.3 seconds for this item with a very small standard deviation among interviewers of approximately 1 second.

Coefficient analysis: P-NonTime model

We extend the above line of analysis for the non-time paradata measures as well, focusing on the 7 items for which $\hat{p}^{P-NonTime}_{Predicted}$ is more than 25%. For clearer display, we show the occurrence proportions in Table 4.5 and the coefficient-related results separately in Table 4.6.

Looking at Table 4.5, we find that 'mean item visits' is the most important non-time paradata measure in terms of occurrence across items, followed by the CV of item visits. We see the presence of 'proportion remarks' for items BCDE1_3 (looking for work, unemployed) and A44 (Telephone/TV/Internet expenses), responses for which were seen to be impacted by DE time in the P-time analysis above. We could surmise that remarks account for at least some of the 'extra' DE time spent by interviewers for these items. We do not find many interactions displayed in Table 4.5; the presence of many competing variables makes it difficult for an interaction to have a occurrence proportion greater

than our minimum display threshold (lowering the threshold to 50% allowed too many interactions reducing display clarity).

We now turn to Table 4.6 which displays the $\hat{\beta}$ (mean and standard deviation) for the non-time measures, and the proportion of resamples in which the coefficient has the same sign. Unlike the time variables, we see that many non-time coefficients have a high standard deviation compared to the mean. This is because many non-time paradata measures are sparse (as we can gauge from Figure 4.3), unlike the time variables which are ubiquitous for every item. In addition, we also have more potentially competing and correlated variables. This coefficient instability is also reflected in the low proportion of models for which the coefficient maintains the same sign, e.g., only approximately 60% of models have a positive sign for the mean item visit measure for item H61e_3. Despite these individual instabilities, the non-time measures, when used together, are generally more powerful than the time measures in explaining $\hat{\sigma}^2_{iwer}$ as we saw in Section 4.5.2.

Table 4.5: Occurrence proportions for the non-time paradata measures. These are proportions of times a non-time measure occurs in the bootstrap models for the 7 items for which $\hat{p}^{P-NonTime}_{Predicted}$ is greater than 25%. Columns are arranged in descending order of $\hat{p}^{P-NonTime}_{Predicted}$. Only main effects with an occurrence proportion greater than 0.8 are shown; the minimum threshold for quadratic terms and interactions is 0.6. The top three main effects in terms of occurrence are in bold. The B, M, and N in the item labels indicate binomial, multinomial, and numeric response types for those items.

	H61e_3 (Medicare insurance. M)	BCDE1_3 (Empl. status - looking for work, unemployed, M)	G12 (Earn wages/salaries apart from uninc. business?, B)	A44 (Telephone/TV/ Internet expenses, N)	A40_1_2 (Home heating - electricity, M)	A4_4 (Dwelling - Mobile home/Trailer, M)	A42_1 (One bill and can separate amounts, M)
$\hat{p}^{P-NonTime}_{Predicted}$	79%	67%	59%	49%	43%	37%	29%
Main effects							
Item visits (mean)	0.88	0.80	**0.86**	**0.91**	0.91	**0.80**	**0.91**
Item visits (CV)	0.87	**0.81**	**0.84**	0.87	0.87		**0.89**
Keycounts (mean)				**0.95**			
Keycounts CV)				**0.93**			
Mouseclicks (mean)	**0.92**		**0.82**	0.86	**0.92**	**0.81**	**0.90**
Mouseclicks (CV)	**0.92**	**0.82**			**0.93**		0.85
Prop. error messages	0.86						
Prop. help	**0.90**			0.84	0.87		0.86
Prop. remarks	0.83	**0.82**		**0.97**	**0.96**		0.85
Quadratic terms							
Mouseclicks (CV)					0.65		
Interactions							
Item visits (CV): Mouseclicks (CV)	0.71						
Item visits (CV) : Prop. help							0.70
Mouseclicks (mean) : Mouseclicks (CV)	0.75						
Mouseclicks (mean) : Prop. help	0.66						0.63

Table 4.6: Coefficient estimates of the non-time paradata variables. These are for the 7 items for which $\hat{p}^{P-NonTime}_{Predicted}$ is greater than 25%. Columns are arranged in descending order of $\hat{p}^{P-NonTime}_{Predicted}$. The B, M, and N in the item labels indicate binomial, multinomial, and numeric response types for those items. The $\hat{\beta}$ correspond to scaled measures used as inputs in the models. All the models for these items, except for A44, were logistic multilevel models. A linear multilevel model was fit for item A44.

	H61e_3 (Medicare insurance. M)		BCDE1_3 (Empl. status - looking for work, unemployed, M)		G12 (Earn wages/salaries apart from uninc. business?, B)		A44 (Telephone/TV/ Internet expenses, N)		A40_1_2 (Home heating - electricity, M)		A4_4 (Dwelling - Mobile home/Trailer, M)		A42_1 (One bill and can separate amounts, M)	
	Mean $\hat{\beta}$ (sd)	Prop. same coef. sign	Mean $\hat{\beta}$ (sd)	Prop. same coef sign	Mean $\hat{\beta}$ (sd)	Prop. same coef sign	Mean $\hat{\beta}$ (sd)	Prop. same sign	Mean $\hat{\beta}$ (sd)	Prop. same sign	Mean $\hat{\beta}$ (sd)	Prop. same sign	Mean $\hat{\beta}$ (sd)	Prop. same sign
Main effects														
Item visits (mean)	0.03 (0.12)	0.59	-0.08 (0.12)	0.76	-0.07 (0.06)	0.89	7.42 (5.25)	0.98	-0.13 (0.06)	0.99	-0.08 (0.15)	0.74	0.07 (0.08)	0.80
Item visits (CV)	0.13 (0.12)	0.84	0.15 (0.14)	0.97	0.05 (0.06)	0.79	-4.79 (6.16)	0.66	0.06 (0.07)	0.77			0.05 (0.08)	0.75
Keycounts (mean)							-7.17 (4.09)	0.97						
Keycounts CV)							17.23 (5.29)	1.00						
Mouseclicks (mean)	-0.2 (0.14)	0.88			-0.1 (0.05)	0.97	-1.62 (2.37)	0.77	-0.05 (0.11)	0.62	0.43 (0.4)	0.99	-0.09 (0.08)	0.91
Mouseclicks (CV)	-0.22 (0.21)	0.81	0.18 (0.17)	0.97					0.03 (0.07)	0.60			-0.03 (0.04)	0.75
Prop. error messages	-0.15 (0.05)	1.00												
Prop. help	0.04 (0.08)	0.69					-6.05 (3.12)	1.00	-0.02 (0.02)	0.78			-0.05 (0.08)	0.74
Prop. remarks	-0.06 (0.05)	0.84	0.15 (0.08)	1.00			-5.13 (4.68)	0.88	-0.19 (0.03)	1.00			0.14 (0.03)	1.00
Quadratic terms														
Mouseclicks (CV)									-0.15 (0.04)	1.00				
Interactions														
Item visits (CV): Mouseclicks (CV)	0.4 (0.22)	1.00												
Item visits (CV) : Prop. help													0.41 (0.2)	1.00
Mouseclicks (mean) : Mouseclicks (CV)	-0.5 (0.11)	1.00												
Mouseclicks (mean) : Prop. help	0.08 (0.07)	0.86											0.17 (0.05)	1.00

Just as we did for the time measures, we predict the impact on substantive responses based on the non-time measures. We first consider the marginal effects of high values (2 standard deviations above the mean) of mean item time, the measure that occurs for all items with a high frequency. This is shown in Figure 4.10). Item A44 has not been shown in this plot for convenience; responses for all other items are on a common log-odds scale. We see that high values of this measure generally lead to a lower log-odds of the response. One plausible mechanism is that error messages appear in other items due to 'positive' responses in these items. Interviewers therefore revisit these items to change the response to 'negative' to eliminate these messages. The impact for item H61e_3 (Medicare insurance) is in the other direction where, on average, an interviewer with high mean field visits is associated with an increased log-odds of a response. For this question, if the respondent says she does *not* have Medicare, a special instruction comes up where the interviewer has to confirm from the respondent that this is the case. If the respondent changes her answer, the interviewer *goes back* to the original item screen and changes the response to this option in the affirmative. Thus, items revisits are likely linked to better probing resulting in a higher log-odds of a positive response.

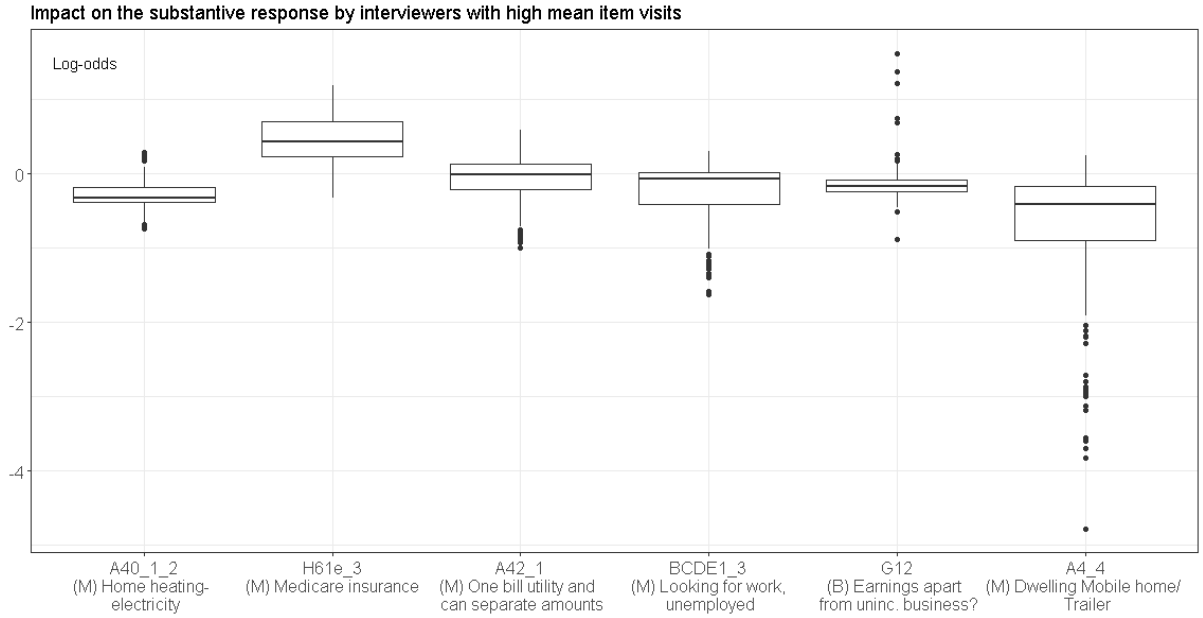

Figure 4.10: Predictions of impact on response for high mean item visits.

Next, in Figure 4.11 we investigate the impact of interactions between help access and remark-making behavior on the response for items H61e_3 (Medicare insurance) and A44 (Telephone/TV/Internet expenses), items for which both measures appear in Table 4.5. For item H61e_3 (Medicare insurance), interviewers who frequently access help for this question (without remarks) are associated with a higher odds of this response option being selected. As explained earlier, this response option has a special check; careful interviewers may be double-checking details by accessing help. The other combinations

do not impact the response. For item A44 (TV/telephone/Internet expense), interviewers with straightforward interviewing (low help, low remarks) are associated with a median $25 higher value. Either one of high remark making or high help access are also associated with a median higher value. If remarks and help are accessed after the first keystroke is entered, then the lower median response value associated with the high remarks-high help combination partially explains the lower median response value for the high DE time combinations seen in Figure 4.9 for this item.

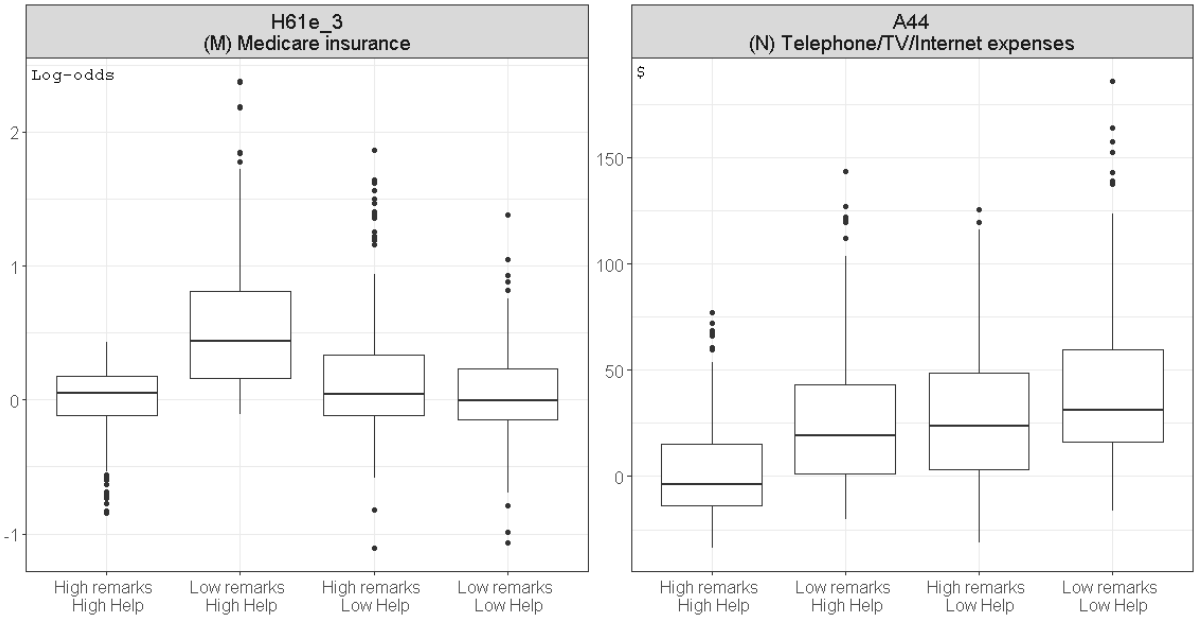

Figure 4.11: Predictions of impact on response by the interaction of help access and remarks. 'High' or 'Low' denotes a proportion that is 2 standard deviations above or below, respectively, from that of the average interviewer.

Comparison of P, NP, and Full models

After exploring the components of the P models, we return to the findings in Section 4.5.2, where the P models did better than the NP models for 5 items, while the NP model did better than the P model for 3 items. In practice, can we then focus on either the NP or P model, as the case may be? To see this, Figure 4.12 plots $\hat{p}_{Predicted}$ for the P and NP models along with the full model. The 11 analysis items can be divided into four groups. The first group has a single member, item A22 (homeowner's insurance premium) where we can use only the NP model given that the P model fails to explain any $\hat{\sigma}_{iwer}$. The second group consists of 3 items where using only the P model is sufficient since the NP variables do not add any incremental explanatory power. These 3 items are: item A44 (monthly telephone/TV/Internet expenses), item A8 (number of rooms), and item H61e_3 (whether Medicare insurance). In fact, for item A44, it might be advisable to use only the paradata variables since the non-paradata variables appear

to be adding more noise in the full model. The third group consists of a single item, BCDE1_3 (looking for work, unemployed) where $\hat{p}_{Predicted}$ is approximately the same for the P and NP models and either model can be used in practice. For the remaining 6 items, the P and NP models are complementary to each other, on average contributing a fairly substantial incremental 0.26 to the larger of the P or NP $\hat{p}_{Predicted}$. These 6 items are: H61J (insurance premium), H61K (was anyone without insurance), A42_1(one single utility bill and can separate amounts), A40_1_2 (home heating - electricity), A4_4 (dwelling - mobile home/trailer), and G12 (whether earned wages/salaries apart from unincorporated business).

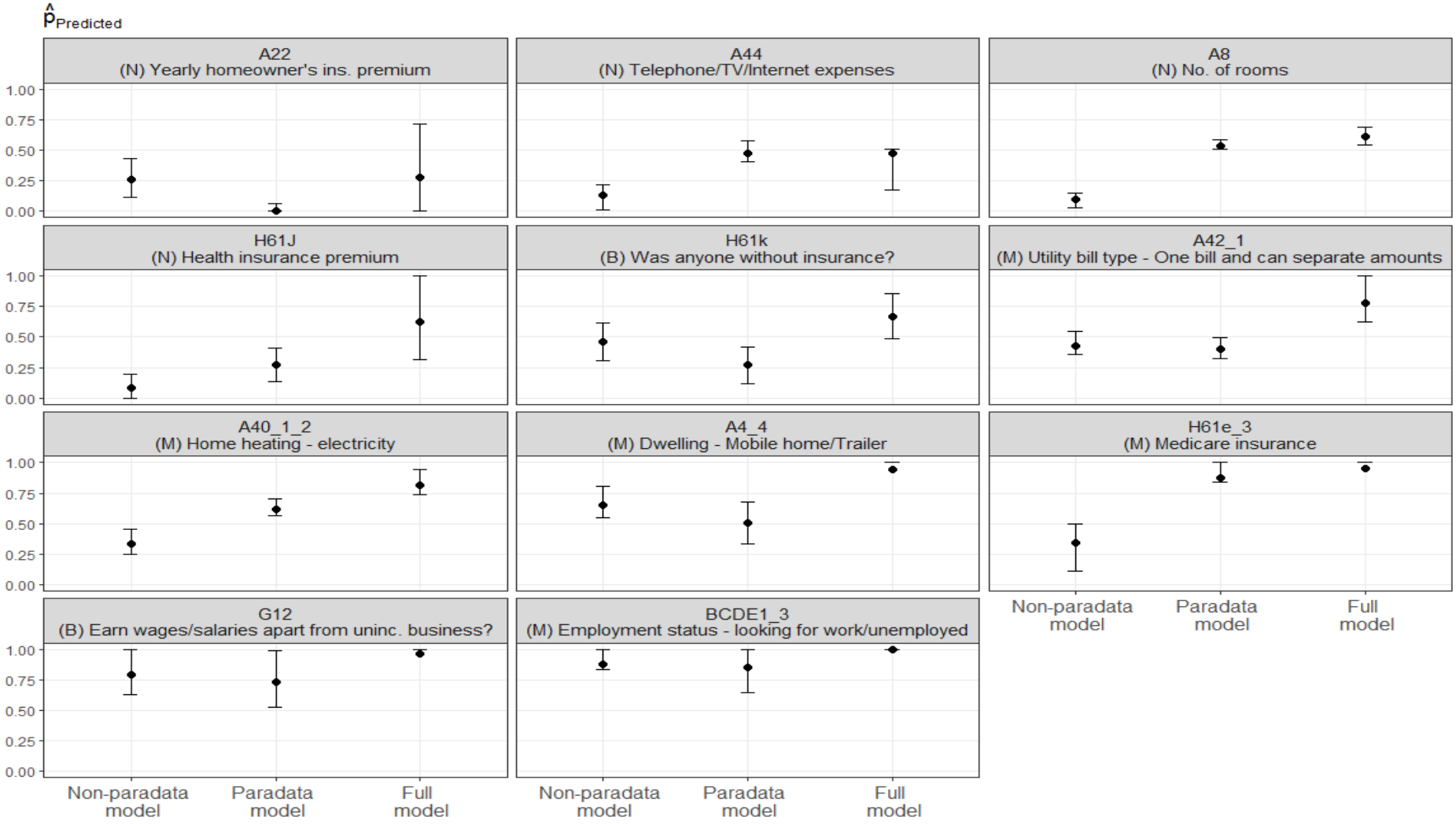

Figure 4.12: Comparison of P, NP, and Full models. Each panel contains estimates of $\hat{p}_{Predicted}$ for the three models along with the IQRs (the lines running through the bars). Panels are arranged in ascending order of $\hat{p}^{Full}_{Predicted}$. The (B), (M), and (N) in the panel labels indicate binomial, multinomial, and numeric response types respectively for the associated items

Figure 4.13 displays diagnostics for the full 'apparent' model (i.e., the model fit to the original data). Except for items A22 (yearly homeowner's insurance premium), A44 (telephone/TV/Internet expenses), and H61J (Health insurance premium) we do not see any major problems in model fit. Of the 3 items with evidence of model fit issues, items A22 and A44 are also those where we see a fairly substantial predictive optimism in Figure 4.6. In such cases, one could try variable transformations to help improve model fit.

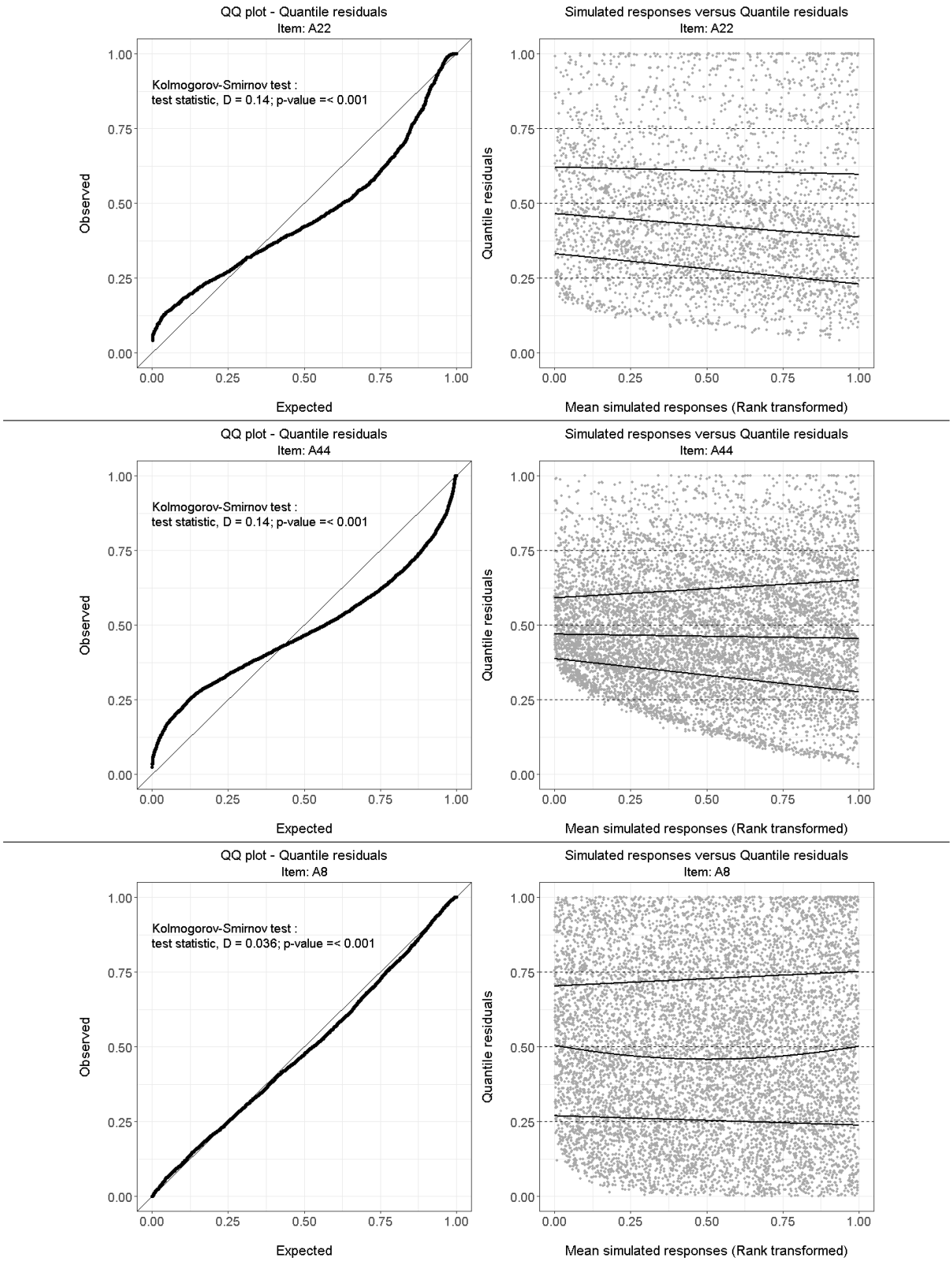

Figure 4.13: Quantile residual diagnostic plots for the full models. The plots are based on the 'apparent' models. Each item is on one row of the plot with the order of the items as followed in Figure 4.12. The left panels compare the quantile residuals to draws from a uniform distribution. Each point in the right panels is the mean simulated response (across 1000 simulations) for an observation in the data. The solid lines in the correspond to the quantile regression lines and the dotted lines are benchmarks for these lines.

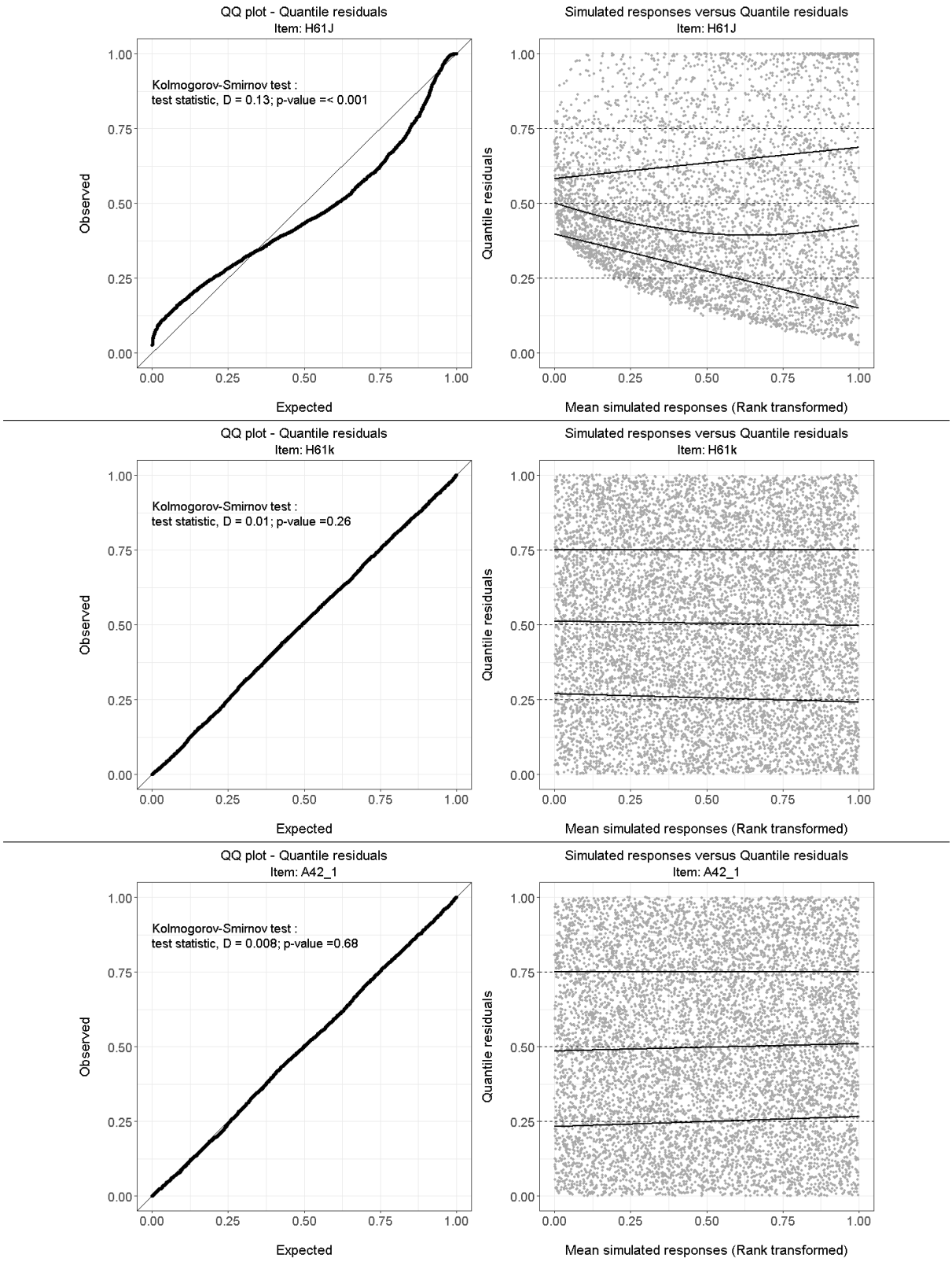

Figure 4.13 (continued).

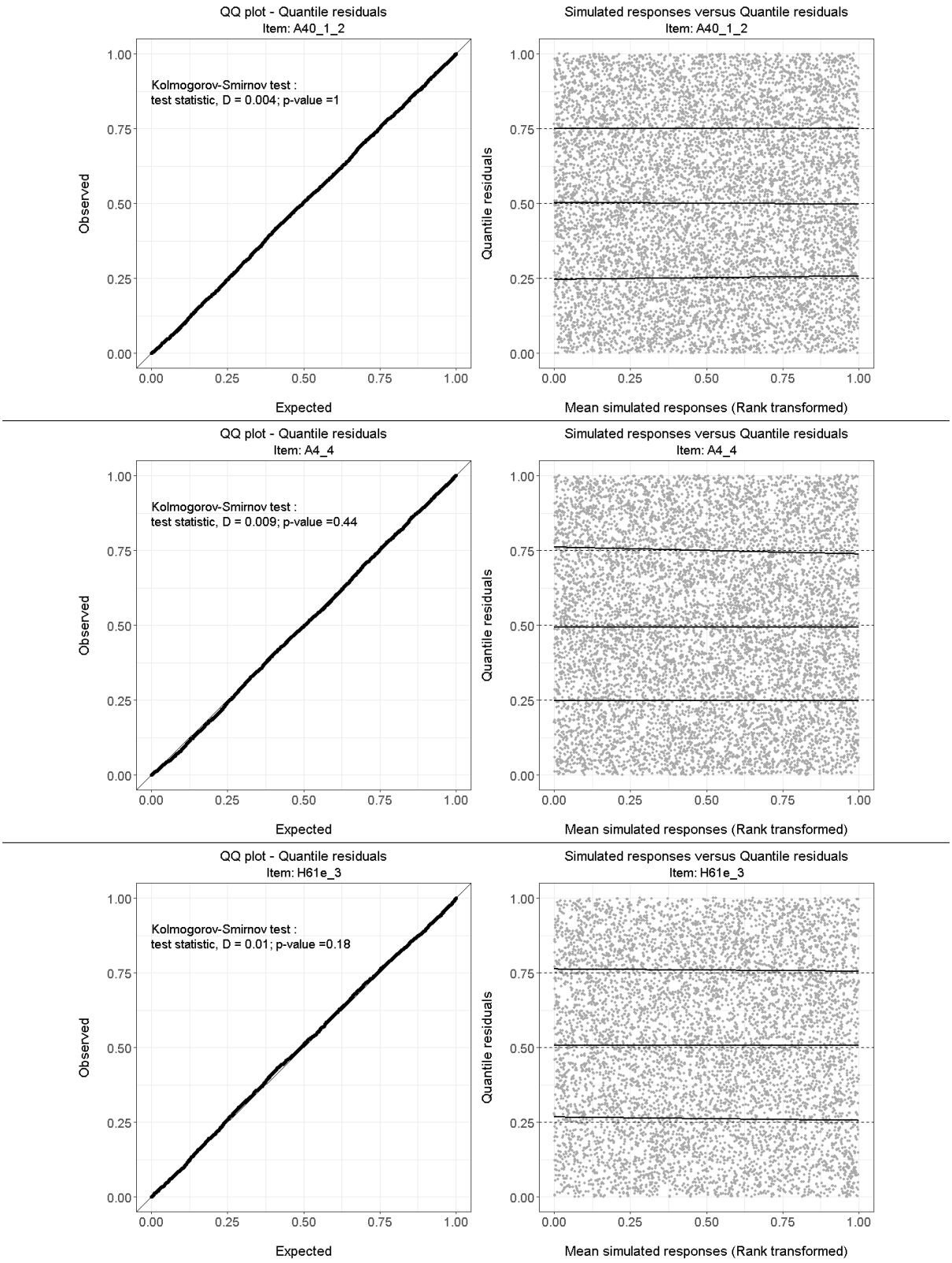

Figure 4.13 (continued).

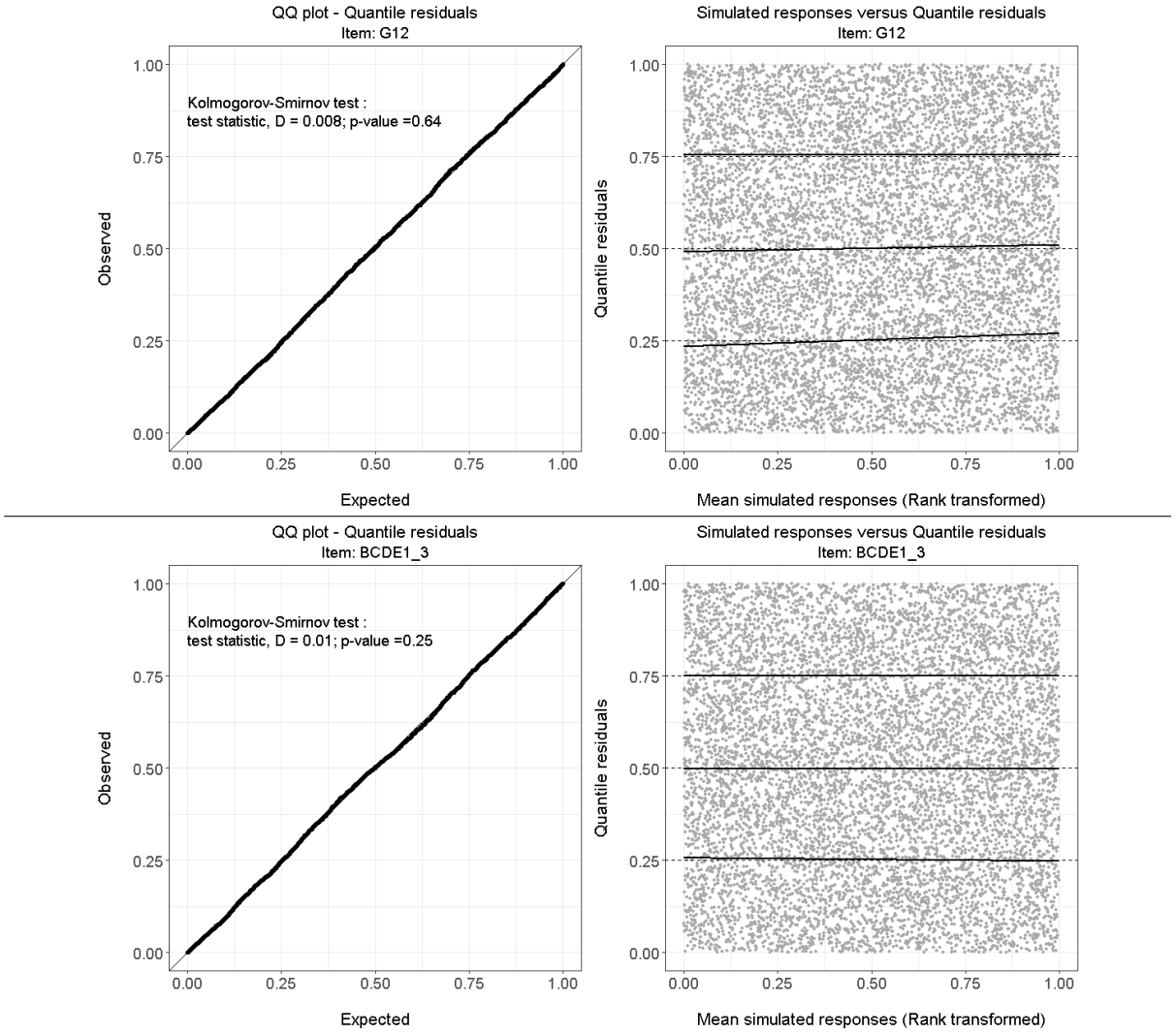

Figure 4.13 (continued).

4.6 Discussion

The results establish that paradata are quite successful in explaining interviewer effects. This success is likely due to paradata being able to capture interviewer behaviors that drive interviewer effects. The impact-on-response analysis showed that paradata are also able to pick up item-specific behaviors that matter. With these encouraging results, a tempering finding was that while the paradata models generally outperformed the non-paradata models, the latter are still important in being able to add incremental explanatory power. This suggests that behaviors or respondent interactions involving, say, different levels of interviewer education are not fully captured by paradata. In practice, it is therefore generally advisable to use both P and NP variables.

The results pertaining to the non-time paradata measures are noteworthy since the paradata literature has focused on item times. The success of the non-time variables could be because they not only capturing information present in the time variables (e.g., remark making behavior is also likely reflected in higher DE times) but different non-time measures are capturing different facets of item-specific interviewer-respondent interactions. This was seen in our results which showed that the frequency of occurrence and impact of different non-paradata measures on the response varies by item. Studies on the lines of Couper and Kreuter (2013) and Olson and Smyth (2015) that have studied item, respondent, and interviewer associations with item times, should also be extended to the non-time paradata variables to better understand the properties of these variables. For the time paradata, the results showed that splitting item times into APR and DE times added more insight. We encourage this split as a standard research practice wherever possible, given that these components are associated with different stages of the interview process involving different types of interviewer and respondent behaviors.

The results showed that the CV variables were important predictors. In the context of APR times, this could be a reflection of inconsistent probing by interviewers, a phenomenon conceived to be arising due to interviewer expectations of the respondents answers (Biemer and Lyberg 2003, p.172). Inconsistent probing has also been posited as a cause of interviewer effects (Biemer and Lyberg 2003; Kreuter 2008), though "so far there has been no systematic study of this effect". Whether the CV variables are specifically getting at this interviewer behavior, and whether this behavior is associated with interviewer effects is an avenue for future work.

Drilling down to the actual measures showed that no single paradata variable is dominant but conversely, there was no paradata measure that did not frequently occur for at least 1 item. Our inferences are conditional on how we defined the measures. It is possible that more creatively defined measures could perform better. To be able to capture as

many behavioral dimensions as possible, we also experimented with other measures, e.g., splitting the error messages into 'hard' errors (which force the interviewer to correct any inconsistency and only then move forward) and 'soft' errors (which the interviewer can simply escape), the proportion of item revisits that involve response edits, whether the interviewer ever exited an interview prematurely at that particular item, etc. However, we found that being sparse, such measures increased predictive uncertainty, a finding echoed in previous literature (Xu et al. 2008; Flynn et al. 2017). Moreover, there was no substantial improvement in average prediction levels when these measures were included. In a similar vein, we caution practitioners from over-relying on automatic variable selection methods to choose the right variables (Belloni et al. 2014, p.40); prior thought, motivated by theoretical findings or practical observations, is important to arrive at the right measures. We checked if we could get more stable predictions (especially for the non-time paradata) if we used only main effect models. However, $\hat{p}_{Predictions}$ were far smaller than what we obtained with the models we presented in this chapter.

Our research in this area is far from comprehensive, and subject to limitations. We used the simple bootstrap method (Efron and Tibshirani 1993) to give us realistic predictions. A better option is to use the enhanced bootstrap (Efron and Tibshirani 1993, p.248-249; Harrell et al. 1996) that adds an additional step of subtracting the optimism indices from the apparent estimates. We did not do this since we obtained 5 zero values of $\hat{p}_{Apparent}$ (Figure 4.4); using the enhanced bootstrap would mean getting a negative value of $\hat{p}_{Prediction}$. However, this means that our predictions may still be optimistic. On the other hand, given our research objectives, our model building did not include interactions between the variable blocks which would potentially add to predictive power. Since, on average, 63% of all interviewers would be in a bootstrap resample (Efron 1983; Efron and Tibshirani 1997), one criticism could be that our predictions are not validated on a fully independent sample. However, as pointed out earlier, approximately 60% of PSID interviewers are duplicated between waves so the bootstrap approach indeed gives us a realistic estimate of predictive power.

While the findings of this research are promising, this research is also subject to several limitations, many of which are avenues for future work. We list 11 such limitations. First, our models rely on the successful approximation to an interpenetrated design but we do not know if we were successful in this endeavor. Second, we are assuming a normal distribution for the random interviewer effects; conceptually, this stands for the effect of many small unobserved interviewer influences so that the central limit theorem would apply. But in our case, we could have many 'average' interviewers and a small number of interviewers who are large exceptions at the tails of the distribution, representing a heavy-tailed distribution. In such cases, there is marked disagreement in the literature as to whether parametric assumptions are important or innocuous (McCulloch and Neuhaus

2011). Some studies such as Maas and Hox (2004) and McCulloch and Neuhaus (2011) show that estimation of the random effect variance is robust to misspecification of the distribution, but one needs to be careful if interest is in the statistical significance of the variance. Grilli and Rampichini (2015) say that "an appropriate specification is crucial for valid predictions of the random effects". Future work could conduct sensitivity analyses on how distributional assumptions impact predictions. Third, we assumed independence of random effects and residual errors. We are not aware of a standard way to adjust our inferences in the event should this assumption not hold in practice. This is another area of future research, including exploring survey mechanisms where this assumption might not hold. Fourth, by looking at a random intercept model, we are looking at the average response. But, if interviewer-specific biases are all in the same direction, we will fail to pick up any interviewer effect. The literature has started looking at this issue (Peytchev 2006; Brunton-Smith et al. 2017) and it would be interesting to see if paradata can explain within-interviewer response variations as well as between-interviewer variability.

Fifth, recent work has shown that responses within interviewers may appear correlated because different interviewers successfully obtain cooperation from different pools of respondents (West and Olson 2010). In our case, this might not matter as much since PSID has a response rate of approximately close to 90% (computed with respect to the previous wave), but future research can consider ways to separately model the measurement error component of the total between-interviewer variance. Sixth, and in a similar vein, models that use paradata to predict interviewer non-response variance would be useful. Seventh, our models ignore the time element; it is possible that associations between interviewer measurement error and paradata patterns change as fieldwork proceeds — and differently for different types of interviewers. Eighth, by conducting an item-level analysis, we ignore correlation of effects between items; interviewer behaviors on an item may impact responses for later items. Ninth, we saw that our models are subject to a fair degree of optimism. One approach to get more stable predictions could be to refit models after using only variables that appear in, say, 50% of resample models. Tenth, while the CV variables were found to be generally important predictors, they could also be unstable especially when used in the early stages of a survey; it is worth assessing the stability of the CV measures. Eleventh, while we gave plausible explanations for our results, going back to recordings where possible and confirming the mechanisms involved will contribute to our knowledge of paradata and interviewer behavior. We did not have access to the recordings and therefore could not undertake this exercise.

4.7 Practical implementation

How do we envisage this research being practically applied? As stated at the outset of this dissertation, and in Section 1 of this chapter, our goal is to develop an interviewer monitoring system that is quick, inexpensive, tailored, and one that does not overly add to the burden on operations' staff. Figure 4.17 shows one way this can occur.

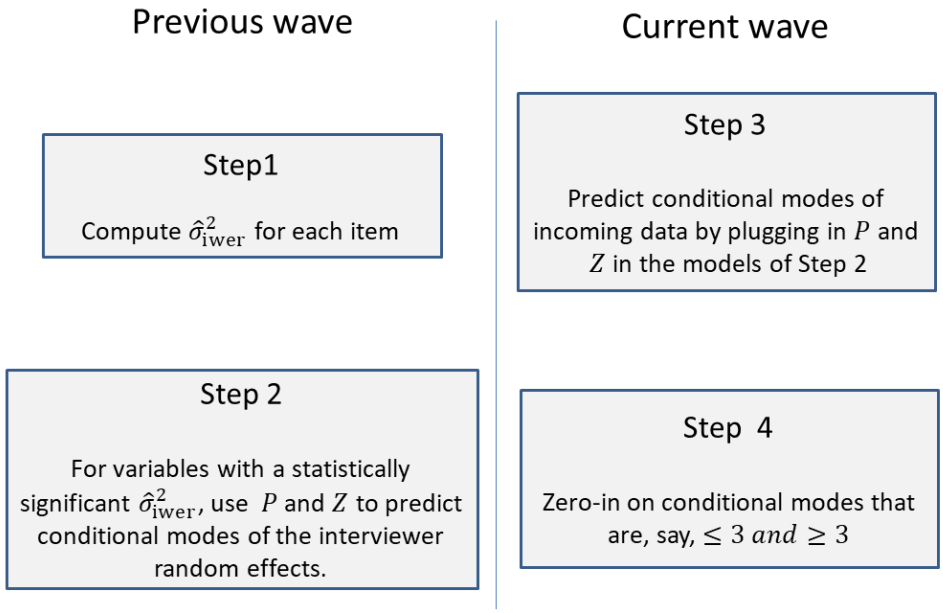

Figure 4.17: Implementation Flow.

Comparing equations 4.2 and 4.3, we see that we are trying to use the vector of paradata and interviewer-level covariates to predict u_{0i}:

$$u_{0i} = P_i^T \boldsymbol{\beta}_P + Z_i^T \boldsymbol{\beta}_Z + u'_{0i} \tag{4.6}$$

As stated in Section 1.3, our focus is on repeated cross-section or panel surveys. Given the nature of these surveys, an overwhelming majority of questionnaire items are constant across waves. Once a wave is completed, σ^2_{iwer} is estimated for all items (Step 1) and for those items with significant σ^2_{iwer}, we fit the models in equations 4.3 or 4.5 using $\mathbf{P}$ and $\mathbf{Z}$ (Step 2). The models are thus kept ready prior to the commencement of a fresh wave. When the current wave starts, incoming $\mathbf{P}$ and $\mathbf{Z}$ are plugged into these models to predict the interviewer conditional modes (Step 3). The crux is that we do not need data on the actual responses to estimate the interviewer effects. This gets around the issue of estimation instability. We also do not need respondent covariate information since these have been controlled for when estimating the model coefficients in equations 4.3 or 4.5.

From this analysis, we will obtain an i x k matrix of predicted conditional modes. We

could then focus on conditional modes towards the extremes of the random effect distribution (Step 4). If recordings are available, they could be reviewed for these specific items and interviewers to check if anything is amiss with the interview. If so, the interviewers in question could be telephoned and be given feedback. In the absence of recordings, more general feedback could be provided to interviewers identified by their conditional modes.

Appendix

4.A Summary descriptive statistics for the items used for the analyses.

Table 4.A.1: Descriptive statistics for the 10 variables used for the analyses. Item descriptions are not precise; please see the questionnaire for the detailed question.

Numeric response variables

Item	Item description	Minimum	Q1	Median	Mean	Q3	Maximum
A8	Number of rooms	0	4	5	5	7	20
A22	Total yearly homeowner's insurance ($)	1	600	950	1,102	1,300	9,000
A44	Telephone and internet expenses ($)	0	90	175	196	270	436
H61J	Monthly health insurance amount ($)	0	75	188	252	350	4,992

Binary response variables

G12	Any salaries or wages besides uninc. business? Yes = 76%
H61K	Any family member without health insurance? Yes = 23%

Categories of Multinomial response variables

A4 (category 4)	Dwelling type - Mobile home/trailer. Proportion = 5%
A40 (category 2)	Heating 1st mention - Electricity. Proportion = 40%
A42 (category 1)	Receive one utility bill and can separate amounts. Proportion = 12%
BCDE1 (category 3)	Employment status - Looking for work, unemployed. Proportion = 7%
H61e (category 3)	Type of health insurance - Medicare. Proportion = 15%

4.B R code for model diagnostics

```
require(DHARMa)
require(qrnn)
require(ggplot2)

#simulate observations
gof_model <- simulateResiduals(model, #name of the model
                               #1000 responses simulated
                               n = 1000,
                               refit = F,
                               #condition on all random effects
                               re.form = NULL)

#mean of the simulated responses for each observation
mean_simresponse <- gof_model$fittedPredictedResponse

#rank transform the mean simulated responses for better visualization
mean_simresponse <- rank(mean_simresponse, ties.method = "average")
mean_simresponse <- mean_simresponse/max(mean_simresponse)

#extract quantile residuals
scaled_resids <- gof_model$scaledResiduals

#data frame for plots
quantresids_data <- data.frame(scaled_resids = scaled_resids, #quantile
                               Expected = runif(gof_model$nObs),
                               mean_simresponse = mean_simresponse)

### Quantile regression
#penalty factor kept as 1 to reduce overfitting
#25th percentile
fit25_nl <- qrnn.fit(x = as.matrix(quantresids_data$mean_simresponse),
                     y = as.matrix(quantresids_data$scaled_resids),
                     n.hidden = 4, iter.max = 1000,
                     n.trials = 1, penalty = 1,
                     tau = 0.25)
quantresids_data$fit25_nl <- qrnn.predict(
  as.matrix(sort(quantresids_data$mean_simresponse)), fit25_nl)

#median
fit50_nl <- qrnn.fit(x = as.matrix(quantresids_data$mean_simresponse),
                     y = as.matrix(quantresids_data$scaled_resids),
                     n.hidden = 4, iter.max = 1000,
                     n.trials = 1, penalty = 1,
                     tau = 0.5)
```

```
quantresids_data$fit50_nl <- qrnn.predict(
  as.matrix(sort(quantresids_data$mean_simresponse)), fit50_nl)

#75th percentile
fit75_nl <- qrnn.fit(x = as.matrix(quantresids_data$mean_simresponse),
                     y = as.matrix(quantresids_data$scaled_resids),
                     n.hidden = 4, iter.max = 1000,
                     n.trials = 1, penalty = 1,
                     tau = 0.75)
quantresids_data$fit75_nl <- qrnn.predict(
  as.matrix(sort(quantresids_data$mean_simresponse)), fit75_nl)

#Kolmogorov-Smirnov test - uniform reference distribution
#used for annotation in the QQ plot
ks.test(quantresids_data$scaled_resids, 'punif')


#QQ plot (theme elements and annotations not shown for brevity)
p_quantresids <- ggplot(quantresids_data,
                        aes(x = sort(Expected),
                            y = sort(scaled_resids))) +
  geom_abline(slope = 1, intercept = 0) +
  ggtitle("QQ plot - Quantile residuals",
  subtitle = "(Interview-level mode)1") +
  xlab("Expected") + ylab("Observed")

#Plot of mean simulated responses against quantile residuals
#(theme elements and annotations not shown for brevity)
p_fitted_quantresids <- ggplot(quantresids_data,
                               aes(x = mean_simresponse,
                               y = scaled_resids)) +

  #quantile regression lines
  geom_line(aes(x = sort(mean_simresponse), y = fit25_nl), size = 1) +
  geom_line(aes(x = sort(mean_simresponse), y = fit50_nl), size = 1) +
  geom_line(aes(x = sort(mean_simresponse), y = fit75_nl), size = 1) +

  #reference lines
  geom_abline(slope = 0, intercept = 0.25, linetype = 2) +
  geom_abline(slope = 0, intercept = 0.50, linetype = 2) +
  geom_abline(slope = 0, intercept = 0.75, linetype = 2) +

  ggtitle("Simulated responses versus Quantile residuals",
  subtitle = "(Interview-level model)") +
  xlab("Mean simulated responses (Rank transformed)") +
  ylab("Quantile residuals")
```

References

Auriat, N. (1993). "My Wife Knows Best": A Comparison of Event Dating Accuracy Between the Wife, the Husband, the Couple, and the Belgium Population Register. *Public Opin. Q.*, 57(2):165.

Austin, P. C. and Merlo, J. (2017). Intermediate and advanced topics in multilevel logistic regression analysis. *Stat. Med.*, 36(20):3257–3277.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.*, 67(1).

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. *J. Econ. Perspect.*, 28(2):29—-50.

Biemer, P. and Lyberg, L. (2003). *Introduction to Survey Quality*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Biemer, P. and Stokes, S. (1985). Optimal Design of Interviewer Variance Experiments in Complex Surveys. *J. Am. Stat. Assoc.*, 80(389):158–166.

Billiet, J. and Loosveldt, G. (1988). Improvement of the Quality of Responses to Factual Survey Questions by Interviewer Training. *Public Opin. Q.*, 52(2).

Brunton-Smith, I., Sturgis, P., and Leckie, G. (2017). Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location-scale model. *J. R. Stat. Soc. Ser. A (Statistics Soc.*, 180(2):551–568.

Cannell, C., Miller, P., and Oksenberg, L. (1981). Research on interviewing techniques. *Sociol. Methodol.*, 12:389–437.

Cannon, A. (2011). Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Comput. Geosci.*, 37(9):1277–1284.

Couper, M. (1998). Measuring survey quality in a CASIC environment. In *Proc. Jt. Stat. Meet. Am. Stat. Assoc. Surv. Res. Methods Sect.*, pages 41–49, Alexandria, VA. American Statistical Association.

Couper, M., Horm, J., and Schlegel, J. (1997). Using trace files to evaluate the National Health Interview Survey CAPI Instrument. In *Proc. Surv. Res. Methods Sect. ASA*, pages 825–829, Anaheim, CA.

Couper, M. and Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *J. R. Stat. Soc. Ser. A (Statistics Soc.*, 176(1):271–286.

Dahlhamer, J., Cynamon, M., Gentleman, J., Piani, A., and Weiler, M. (2010). Minimizing Survey Error through Interviewer Training : New Procedures Applied to the National Health Interview Survey (NHIS). In *Proc. Jt. Stat. Meet. Sect. Surv. Res. Methods*, pages 4627–40. American Statistical Association.

Dunn, P. and Smyth, G. (1996). Randomized Quantile Residuals. *J. Comput. Graph. Stat.*, 5(3):236.

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *J. Am. Stat. Assoc.*, 78(382):316–331.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Springer US, Boston, MA.

Efron, B. and Tibshirani, R. (1997). Improvements on Cross-Validation: The .632+ Bootstrap Method. *J. Am. Stat. Assoc.*, 92(438):548.

Fellegi, I. (1974). An Improved Method of Estimating the Correlated Response Variance. *J. Am. Stat. Assoc.*, 69(346).

Flynn, C., Hurvich, C., and Simonoff, J. (2017). On the Sensitivity of the Lasso to the Number of Predictor Variables. *Stat. Sci.*, 32(1):88–105.

Fowler, F. and Mangione, T. (1990). *Standardized Survey Interviewing: Minimizing Interviewer Related Error*. SAGE Publications, Inc., Thousand Oaks, California.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Grilli, L. and Rampichini, C. (2015). Specification of random effects in multilevel models: a review. *Qual. Quant.*, 49(3):967–976.

Groves, R. (1989). *Survey Errors and Survey Costs*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Groves, R. and Couper, M. (1998). *Nonresponse in Household Interview Surveys*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Groves, R., Fowler, F. J., Couper, M., Lepkowski, J., Singer, E., and Tourangeau, R. (2004). *Survey Methodology*. Wiley, New York.

Hansen, M., Hurwitz, W., and Bershad, M. (1960). Measurement Errors in Censuses and Surveys. *Bull. Int. Stat. Inst.*, 32(2):359–374.

Harrell, F., Lee, K., and MARK, D. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.*, 15(4):361–387.

Hartig, F. (2018). DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. `https://cran.r-project.org/package=DHARMa`.

Hicks, W., Edwards, B., Tourangeau, K., McBride, B., Harris-Kojetin, L., and Moss, A. (2010). Using Cari Tools To Understand Measurement Error. *Public Opin. Q.*, 74(5):985–1003.

Hox, J. (2010). *Multilevel Analysis: Techniques and Applications*. Routledge, New York, NY, second edition.

Jans, M., Sirkis, R., Schultheis, C., Gindi, R., and Dahlhamer, J. (2011). Comparing CAPI Trace File Data and Quality Control Reinterview Data as Methods of Maintaining Data Quality. In *Proc. Jt. Stat. Meet. Sect. Surv. Res. Methods*, pages 404–17. American Statistical Association.

Johnson, S. The NLopt nonlinear-optimization package.

Kenkel, B. and Signorino, C. (2018). polywog: Bootstrapped Basis Regression with Oracle Model Selection.

Kish, L. (1962). Studies of Interviewer Variance for Attitudinal Variables. *J. Am. Stat. Assoc.*, 57(297):92.

Knauper, B. (1999). The Impact of Age and Education on Response Order Effects in Attitude Measurement. *Public Opin. Q.*, 63(3):347.

Kreuter, F. (2008). Interviewer Variance. In Lavrakas, P., editor, *Encycl. Surv. Res. Methods*. Sage Publications, Inc., Thousand Oaks, CA.

Krosnick, J. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Appl. Cogn. Psychol.*, 5(3):213–236.

Krosnick, J. and Alwin, D. (1987). An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opin. Q.*, 51(2):201.

Lee, J. and Lee, S. (2012). Does it Matter WHO Responded to the Survey? Trends in the U.S. Gender Earnings Gap Revisited. *ILR Rev.*, 65(1):148–160.

Maas, C. and Hox, J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Comput. Stat. Data Anal.*, 46(3):427–440.

Mahalanobis, P. (1946). Recent Experiments in Statistical Sampling in the Indian Statistical Institute. *J. R. Stat. Soc.*, 109:325–370.

McCulloch, C. and Neuhaus, J. (2011). Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter. *Stat. Sci.*, 26(3):388–402.

Nicolaas, G. (2011). ESRC National Centre for Research Methods Review paper Survey Paradata : A review. Technical Report January, ESRC National Centre for Research Methods.

Olson, K. and Smyth, J. (2015). The Effect of CATI Questions, Respondents, and Interviewers on Response Time. *J. Surv. Stat. Methodol.*, 3(3):361–396.

Peytchev, A. (2006). Estimation of Measurement Error and Identification of Causes: Linking Measurement Error to Nonresponse, Interviewers, and Interviewer Characteristics. In *Proc. Jt. Stat. Meet. Surv. Res. Methods Sect.*, pages 3528–3535. American Statistical Association.

Schnell, R. and Kreuter, F. (2005). Separating Interviewer and Sampling-Point Effects.

J. Off. Stat., 21(3):389–410.

Self, S. and Liang, K.-Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. *J. Am. Stat. Assoc.*, 82(398):605–610.

Skowronski, J. and Thompson, C. (1990). Reconstructing the dates of personal events: Gender differences in accuracy. *Appl. Cogn. Psychol.*, 4(5):371–381.

Snijders, T. and Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage Publishers, London, 1st edition.

Steinkamp, S. (1964). The Identification of Effective Interviewers. *J. Am. Stat. Assoc.*, 59(308):1165–1174.

Stock, S. and Hochstim, J. (1951). A Method of Measuring Interviewer Variability. *Public Opin. Q.*, 15(2).

Team, R. C. (2013). R: A Language and Environment for Statistical Computing.

Tourangeau, R., Rips, L., and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press, Cambridge.

van der Zouwen, J. and Dijkstra, W. (1988). Types of Inadequate Interviewer Behaviour in Survey Interviews; Their Causes and Effects 1. *Bull. Sociol. Methodol. Méthodologie Sociol.*, 18(1):5–20.

West, B., Kreuter, F., and Jaenichen, U. (2013). Interviewer Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error, or Nonresponse? *J. Off. Stat.*, 29(2).

West, B. and Olson, K. (2010). How Much of Interviewer Variance is Really Nonresponse Error Variance? *Public Opin. Q.*, 74(5):1004–1026.

West, B. and Sinibaldi, J. (2013). The Quality of Paradata: A Literature Review. In *Improv. Surv. with Parad.*, pages 339–359. John Wiley & Sons, Inc., Hoboken, New Jersey.

Xu, H., Shie, M., and Constantine, C. (2008). Sparse algorithms are not stable: A no-free-lunch theorem. In *2008 46th Annu. Allert. Conf. Commun. Control. Comput.*, pages 1299–1303. IEEE.

Ypma, J., Borchers, H., and Eddelbuettel, D. (2014). nloptr - R interface to NLopt.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *J. Am. Stat. Assoc.*, 101(476):1418–1429.

Chapter 5

Conclusion

In this chapter, we review the goals of this dissertation, discuss some key implications of the results, and suggest next steps for future research.

5.1 Review of dissertation goals

The central goal we set at the outset was to explore if paradata could be used in a systematic manner to create an early warning system to spot interviewers likely to be contributing to interviewer effects. We started with the hypothesis that different levels of interviewing quality cause different paradata patterns. Differing levels of interviewing quality also result in different between-interviewer response means even after controlling for respondent characteristics, thus leading to interviewer effects. Thus, interviewing quality was conceptualized as a common cause of both interviewer effects and paradata patterns, making it possible for us to think about directly using paradata to predict interviewer effects. Chapters 2-4 tested if there was evidence to support these hypotheses using the following six streams of data from the 2015 wave of the Panel Study of Income Dynamics (PSID): paradata, interviewing quality data, the early release version of substantive data, respondent characteristics, interviewer characteristics, and item characteristics.

5.2 Key implications of the dissertation results

5.2.1 Paradata and interviewer effects

The results from Chapter 4 provide strong evidence to establish the utility of paradata in predicting interviewer effects. A key strength of the approach in Chapter 4 lies in the manner in which paradata and estimates of interviewer effects are working in tandem. On the one hand, paradata are produced almost free of charge, are relatively error-free, and are available on all cases but need objective outcomes to link with to maximize their utility; interviewer effect estimates provide these links. On the other hand, interviewer effects that are often neglected (Elliott and West 2015) are now in active consideration via their paradata proxies; this research attempts a shift in the thinking that "the assessment of interviewer error is a post-survey quality measure" (Biemer and Lyberg 2003, p.168). From a survey management perspective, this is a step towards resolving the issue that "most methodologists are several degrees distant from operations managers" (Edwards et al. 2017, p.269). Another feature of our methods is that the interviewer quality control (QC) process is now directly tied to substantive responses.

One aspect we did not cover in this dissertation, but is certainly worth considering for future research, is the use of Bayesian methods to update models based on incoming data. Future research should also consider methods for one-time surveys; this dissertation only considered applications to repeated cross-sectional or panel surveys.

5.2.2 What do we mean by interviewing quality?

Results in Chapters 2 showed strong associations between paradata and indicators of interviewing quality, and results in Chapter 4 showed strong associations between paradata and interviewer effects. However, we saw only moderately strong associations between QC flag variables and interviewer effects in Chapter 3. How do we explain this seeming discrepancy? First, we recognize the extensive literature that establishes that differences in interviewing quality among interviewers do cause interviewer effects (e.g., Kish 1962; Groves 1989, West and Blom 2016). Then, the lack of strong effects in Chapter 3 could be due to the shortcomings of the QC variables, as explained in Section 7 of that chapter.

However, consider the case where there are no problems with the QC variables as well. An important aspect to then consider is the quality assurance (QA) process that deals with the interviewing standards to be met - in contrast to the actual implementation of those standards that quality control (QC) deals with. In Chapter 3, an implicit assumption

is that the QC process is being implemented as designed. Then, a lack of variance explanation by the flag variables implies that the interviewing protocols themselves, i.e., the QA process itself, could be ineffective at achieving the desired objective which is to control interviewer effects. If this is the case, perhaps more thought needs to be given to the design of interviewing protocols. For example, rather than sticking to a strictly standardized format, interviewers could be given more flexibility in interviewing for questions that have complex mappings to respondent situations (Schober and Conrad 1997; Schober et al. 2004). However, any such redesign should be preceded by resolving, to the extent possible, issues with the QC flag variables (pointed out earlier) and an evaluation of the robustness of the QC coding process. It would also be advantageous to try to implement an approximately interpenetrated design (Biemer and Lyberg 2003, p.166) to eliminate possible issues with model approximation to interpenetration that lead us to our inferences.

A related idea is that more thought needs to be given to the meaning of 'interviewing quality'. The flag variables used in Chapter 3 were based on interviewer behaviors such as probing, asking questions, etc., which certainly have firm grounding in the literature (Groves 1989; Fowler and Mangione 1990; West and Blom 2016). However, more research is needed on the impact of other phenomena such as interviewer paralinguistic behaviors (Conrad et al. 2008) and interviewer-respondent interactions on interviewer effects; some estimates suggest that half of everything the interviewer says in an interview is other than a question or a probe (Cannell et al. 1968; Fowler and Mangione 1990, p.68). Previous research (Conrad et al. 2013) has shown that moderate use of fillers (e.g., 'um' and 'uh') by interviewers is associated with higher participation rates while simultaneous speech between interviewers and respondents produced more refusals. We could imagine these results having analogues in measurement error as well; hypothetically, the non-use of fillers could communicate a robotic communication style which may discourage respondents from thinking actively thus providing less accurate responses.

Some of the success of paradata over the interviewing quality variables in predicting interviewer effects could be stemming from paradata absorbing some of information present in paralinguistic features. For example, the use of fillers would lead to an increased APR time while simultaneous speech would lead to lower APR times, all else being equal. Anxious interviewing could be reflected in frequent remark-making but also manifest in interviewers' pitch which is perceived by the respondent. Paradata are therefore, we think, not only able to capture behaviors that the QC system does (Chapter 2 results) but they also go beyond this by capturing other information that impact interviewer effects. This, coupled with the strength of paradata such as its flexibility, leads to the strong results in Chapter 4.

Some recent work (Nuttirudee 2015) has looked at associations between interviewer voice features and data quality. But the indicators of data quality used were indirect in nature, e.g., rounding, directional hypotheses such as 'more is better', and respondent behaviors such as interrupting questions with an answer. Future research can consider analyses such as those done in this dissertation - using interviewer effects as direct estimates of data quality. Results of such research can have important consequences for training such as emphasizing that interviewers should not speak over respondents or to modulate ones pitch effectively.

5.2.3 Implications on survey operations

The evidence presented in this dissertation bolsters the case for increased use of paradata in practice, especially given that we tried to be as realistic as possible in our inferences by adopting methods such as adjusting for multiple comparisons in Chapter 2 or undertaking the bootstrap-based predictions in Chapter 4. Survey organizations should seriously consider investing in necessary infrastructure for the parsing, storing, and easy access to paradata. Paradata, as stored by the Computer-Assisted Interviewing (CAI) instrument, is not usable unless parsed by external software. While this will incur some initial software development cost, it is likely to be more or less a fixed cost. Paradata should not be fully processed into measures by such software but left flexible for researchers to define, at least in the initial phase when efficacy of several measures should be tested. This will also allow researchers to examine the incremental value of a particular paradata measure. All this will require close coordination between the operations staff (who typically have strong insights on likely mechanisms producing the paradata) and research staff (whose expertise lies in model building). A strong feedback loop from operations to research is needed when models are being tested on live data so that models can be refined.

Databases should be structured to also incorporate pre-edit substantive data (to enable computing of interviewer effects), QC data (to check interviewing quality status), interviewer characteristics, and respondent characteristics (to use as controls in modeling). Having such a common database across analysts reduces the chances of analytic errors. For implementation, managers can start with a small set of items that are most valuable to the user community and then expand to the full set for modeling, if needed.

5.2.4 Sampling interviews for quality control

Organizations employing Computer-Assisted Recorded Interviewing (CARI) often choose a small number of interviews in advance to be recorded (PSID uses a sampling rate of

approximately 6%). However, it is possible for interviewers to gauge whether an interview is being recorded and improve their performance for only those interviews (McGonagle et al. 2015). This being the case, our recommended strategy is to record all interviews and use the methods outlined in Chapters 2 and 4 to listen to the sampled recordings so that there is concrete evidence of interviewers deviating from protocol. However, there are 2 downsides to this approach. First, not every respondent will consent to the recording (e.g., PSID's consent rate is 94%, McGonagle et al. 2015) so we risk reducing sample size. Second, McGonagle et al. (2015) found that interviewers take an additional 7% longer to conduct an interview when they know that interviews are being recorded. While this incremental time is not very high, there is some risk of increased interviewer fatigue leading to poorer interviewing quality especially towards the later stages of fieldwork.

For surveys such as PSID that have historical QC data, it is advantageous to use the methods in Chapters 2 and Chapter 4 in combination. The interviewer-level predictions based on the models in Chapter 4 are conducted only after a certain proportion of the fieldwork is complete. Until then, the models in Chapter 2 can be used to predict interview-level quality issues; results in Chapter 2 showed that the first 2 interviews have a higher odds of a QC flag thereby illustrating the importance of identifying issues early. Interviews or items-within-interviews with low predicted odds of a flag should also be sampled to give an opportunity to identify positive interviewer behaviors and reinforce these (Couper et al. 1992; Biemer and Lyberg 2003, p. 179).

Even after the Chapter 4 models start getting implemented, Chapter 2 models can be used to zero-in on the exact interview within an interviewer's workload that has potential problems (often needed by supervisors to give concrete feedback). This is especially useful at later stages of fieldwork when the number of completed interviews is high; it would be time-consuming to listen to all interviews of an interviewer even for a single item. Note that we do not rely only on Chapter 2 models since they do not explicitly involve the substantive response values. Another advantage of using Chapter 2 models along with Chapter 4 models arises in situations where an interviewer exhibits inconsistent interviewing across interviews so that her mean response is close to the overall average. Chapter 4 models would miss such interviewers but Chapter 2 models would still identify potentially problematic interviews.

5.3 Future research

While we suggested avenues for future research in each chapter as well as in the preceding discussion, we feel the following 4 areas should receive priority:

1. Replication.
   Given that this is the first study of its kind, replication of this research will add to evidence on hand. The PSID is a survey focusing mainly on questions of economic interest. Replication on surveys that deal with other subject matters or questions of a different type (e.g., attitudinal questions) will be helpful. It would also be useful to see how our results compare to surveys using the face-to-face mode. For this, it is necessary that interviewers not be restricted to a single geographic unit or else geographic and interviewers effects would be confounded. We would also require geographic covariates to control for the non-random assignment of interviewers to geographies (West et al. 2013).

2. Exploring the properties of non-time paradata.
   Studying the properties of non-time paradata is important given its association with interviewer effects as seen in Chapter 4. Couper and Kreuter (2013) look at item times as a function of item characteristics, respondent characteristics and interviewer characteristics for a face-to-face survey; Olson and Smyth (2015) do the same for a telephone survey. Similar studies should be conducted for the non-time variables too. One challenge is that many non-time variables are sparse as we saw in Chapter 4. To overcome this, the PCA approach used in Chapter 2 can be used. Here, the PC scores would be the outcome variable and item, respondent, and interviewer characteristics would be the inputs. Results of this study can have questionnaire design and interviewer training implications, e.g., if we know that certain types of interviewers are more prone to remark making (which may disrupt interview flow), this can be addressed during training.

3. Understanding the mechanisms that drive paradata.
   In Chapters 2 and 4, we advanced likely mechanisms that explained our results involving paradata. For example, we hypothesized that higher mean DE times could mean back-forth with the respondent even after the response is recorded. But these post-hoc explanations can be confirmed only by actually listening to the relevant recordings. This will help map paradata to actual interviewer activities and understand the paradata-generating mechanisms (which would also vary by item type). Another example is that of item revisits which we found to be an important paradata measure in predicting interviewer effects in Chapter 4. We can partially understand why revisits take place by refining this measure as say, 'revisits that also involve response edits'. But this will still not directly answer questions such as 'Are item revisits taking place to compensate for lack of initial probing?'. Undertaking such analysis, even of a qualitative kind, will add to our understanding of paradata and help refine question construction and training efforts.

4. Understanding the role of paralinguistic features in explaining interviewer effects as explained in Section 5.2.2.

Writing in 1981, Cannell et al. say: "Despite the potential for interviewers to bias data somehow, concern over interviewer effects has lessened in recent years". They attributed this to 'improvements in survey practice'. However, close to 4 decades later, the need to focus on interviewer effects is probably more than ever as "longer interviews (are) being requested by interviewers with less experience from persons who are more sensitive to the burden of the request" (Groves 2003). With declining response rates, growing privacy concerns, and respondent reluctance, a lot of focus in recent years has been on non-response, possibly at the cost of measurement error. Part of the reason is also that non-response is 'visible' while measurement error seems to lurk in the background; while response rate targets are set, it is difficult to set a 'data quality' target. We like to imagine survey quality dashboards in the near future having measurement error indicators for interviewers on the lines of 'contribution to ρ_{int}' along with existing non-response and productivity indicators.

This dissertation was designed with a strong emphasis on improving survey practice. We will be rewarded if these ideas find their way into regular survey operations.

References

Biemer, P. and Lyberg, L. (2003). *Introduction to Survey Quality*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Cannell, C., Fowler, F., and Marquis, K. (1968). The Influence of Interviewer and Respondent Psychological and Behavioral Variables on the Reporting of Household Interviews. *Vital Heal. Stat.*, Series 2(No. 26.).

Cannell, C., Miller, P., and Oksenberg, L. (1981). Research on interviewing techniques. *Sociol. Methodol.*, 12:389–437.

Conrad, F., Broome, J., Benkí, J., Kreuter, F., Groves, R., Vannette, D., and McClain, C. (2013). Interviewer speech and the success of survey invitations. *J. R. Stat. Soc. Ser. A (Statistics Soc.*, 176(1):191–210.

Conrad, F., Schober, M., and Dijkstra, W. (2008). Cues of Communication Difficulty in Telephone Interviews. In *Adv. Teleph. Surv. Methodol.*, pages 212–230. John Wiley & Sons, Hoboken, New Jersey.

Couper, M., Holland, L., and Groves, R. (1992). Developing systematic procedures for monitoring in a centralized telephone facility. *J. Off. Stat.*, 8(1):63–76.

Couper, M. and Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *J. R. Stat. Soc. Ser. A (Statistics Soc.*, 176(1):271–286.

Edwards, B., Maitland, A., and Connor, S. (2017). Measurement Error in Survey Operations Management. In *Total Surv. Error Pract.*, pages 253–277. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Elliott, M. and West, B. (2015). Clustering by Interviewer: A Source of Variance That Is Unaccounted for in Single-Stage Health Surveys. *Am. J. Epidemiol.*, 182(2):118–126.

Fowler, F. and Mangione, T. (1990). *Standardized Survey Interviewing: Minimizing Interviewer Related Error*. SAGE Publications, Inc., Thousand Oaks, California.

Groves, R. (1989). *Survey Errors and Survey Costs*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Groves, R. (2003). Trends in Survey Costs and Key Research Needs in Survey Nonresponse. In Tourangeau, R., editor, *Natl. Sci. Found. Work. Recurring Surv. Issues Oppor.*

Kish, L. (1962). Studies of Interviewer Variance for Attitudinal Variables. *J. Am. Stat.*

Assoc., 57(297):92.

McGonagle, K., Brown, C., and Schoeni, R. (2015). The Effects of Respondents' Consent to Be Recorded on Interview Length and Data Quality in a National Panel Study. *Field methods*, 27(4):373–390.

Nuttirudee, C. (2015). *Interviewer Voice Characteristics and DataQuality*. PhD thesis, University of Nebraska-Lincoln.

Olson, K. and Smyth, J. (2015). The Effect of CATI Questions, Respondents, and Interviewers on Response Time. *J. Surv. Stat. Methodol.*, 3(3):361–396.

Schober, M. and Conrad, F. (1997). Does Conversational Interviewing Reduce Survey Measurement Error? *Public Opin. Q.*, 61(4):576.

Schober, M., Conrad, F., and Fricker, S. (2004). Misunderstanding standardized language in research interviews. *Appl. Cogn. Psychol.*, 18(2):169–188.

West, B. and Blom, A. (2016). Explaining Interviewer Effects: A Research Synthesis. *J. Surv. Stat. Methodol.*

West, B., Kreuter, F., and Jaenichen, U. (2013). Interviewer Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error, or Nonresponse? *J. Off. Stat.*, 29(2).