Using Audit Trails to understand the response time in China Family Panel Studies (CFPS)

Yan Sun, Institute of Social Science Survey, Peking University

This paper examines the determinants of response times (RT) to survey questions. Using the unweighted multilevel model, we include item-level characteristics. respondent-level characteristics and interviewer-level characteristics thought to affect response time in a three-level cross-classified model.

The results suggest that response times are affected by question characteristics such as the length of the question and question type that probably reflect reading times and cognitive process. In addition, response times are also affected the location of the question within the questionnaire. Our results replicate findings on respondent characteristics and response times that have been found in prior studies, such as age and education. We found that CFPS interviewers appear to contribute independently to the response times, and their measured demographic characteristics and experience explain a considerable fraction of the interviewer variance component.

Key words: Paradata, item response time, respondent, interviewer

1. Introduction

Paradata are automated data generated directly by the survey data collection process and can be used to describe and evaluate that process (Couper, 2000), which have been widely used nowadays. However, little attention has been paid to audit trails data. Audit trails data are one form of paradata that are generated by computer-assisted interviewing (CAI) instruments. Audit trails data, also called keystroke data, record all keys pressed and/or mouse movement and clicks as interviewers (or respondents) move through a survey. An audit trails file includes both what and when actions occur during an interview. Therefore, it can be used to restore the exact process of how interviewers interact with an instrument. To date, response times have been treated either as a predictor or as a proxy measure for some other variable in most of the studies (Yan and Tourangeau, 2008). Only few studies used response times as the main focus. In this paper different we take on a different perspective on item response time. We considered the response time as a dependent variable, try to explore the determinants of response time.

Why we are interesting in response time? During the field production we noticed that total interview length for each interview dropped significantly as the time being for each interviewer. We have a hypothesis that interviewers might take shortcut by "making" the interview enter fewer modules or ask fewer questions. However, when we look at the number of questions been asked for each interview we did not find any declining pattern, so we believe interviewers do not mean to make the interview shorter. With given evidence, we checked the response time for each question. We found that the shorter interview length is resulted from the shorter response time used per question. According to the fact we found, it is good reason to assume that interviewers do have influence on response times. The aim of this paper is to test whether there is systematic variation in response time by interviewer characteristics. To the extent that this is the case, response time has the potential for signaling interviewer performance.

2. Survey response time

2.1 How to measure response time

Two approaches have been used in measuring response times in the survey literature: active timers and latent timers.

(1) Active response time: the elapse between the interviewer finish reading a question to the respondent start to give an answer.

(2) Latent Response time: the moment the question appear on the interviewer’s monitor to the interviewer coding the answer.

The two different response time lies in the different assumption about when the survey response process starts. Active timers assume that the response process begins only after the question has been completely presented to the respondent. Latent timers assume that the response process begins as soon as the question starts being presented to the respondent.

According to our experience, it is common that respondents interrupt and give an answer before interviewers finish question reading. In addition, empirical evidence has demonstrated that response times obtained via active timers and latent timers are significantly correlated; they produce consistent and comparable model parameter estimates given correct model specification (Mulligan et al., 2003). Since we believe that much of the response process is likely to happen during the interviewer reading the question, the latent timer approach would seem to be more appropriate for our study.

2.2 Determinants of response times

Under interviewer administered mode, there are three parties to an interview—the interviewer, the respondent, and the task, that is, the instrument. Each of these components affect the processes the interview and the length of the response time.

Previous studies have identified that the characteristics of the field effect item level response time. The length of the question has positive effect on response time (Yan and Tourangeau, 2008, Couper and Kreuter, 2013) and the complexity of the question increase the response time. Different question types result in different response time. Question in form of single choice or integer need less the response time than open ending question (Couper and Kreuter, 2013).

With regard to respondent characteristics, age and education is repeatedly found to be related to response time. Age and education are two important factors in determining respondents′ working capacity on answering the questions (see Salthouse (1991) and Schwarz *et al.* (1999), which result in different response time. With reduced times for more educated people (Salthouse, 1991; Yan and Tourangeau, 2008, Yan and Tourangeau (2008) and Couper and Kreuter(2013)). There is no stable finding about the respondent gender on response times.

Little is known about interviewer effects on response time measures, in part, because a large portion of the response time research has been conducted in self-administered modes(e.g. web survey) with no interviewer present or measures of response time were focused on the respondent only. A recent study by Couper and Kreuter found that interviewers are having relatively little influence on completion time (Couper and Kreuter, 2013).

3. Source of data and variables

The data used in this paper are from 2011 survey of Chinese Family Panel Studies (CFPS) conducted by Institute of Social Science Survey, Peking University. The survey is based on a national probability sample of household in 25 province, cities and autonomous regions in China (excluding Hong Kong, Macau, Taiwan and Xinjiang Uyghur Autonomous Region, Tibet Autonomous Region, Qinghai Province, Inner Mongolia Autonomous Region, Ningxia Hui Autonomous Region and Hainan Province). The baseline survey was conducted in 2010. In 2011, a follow-up survey to all the households interviewed in the baseline survey was conducted. In each household, a family financial questionnaire about income and expenditure need to be filled. All the interviews are conducted by CAPI mode. For detailed information, please refer to the official website of Institute of Social Science Survey, Peking University http://www.isss.edu.cn/.

The family financial questionnaire contained up to 7 different modules, including housing, non-agriculture activity, agriculture activity, income and expenditure and so on. The average total interview length was 27 min. The average number of question in an interview is 141. This paper focus on 26 key variables selected for interviewer performance evaluation.

3.1. Description of variables and sources of data

We merged data from four different sources:

(a) the response time, extracted from the Blaise audit trails,

(b) characteristics of the items, created by questionnaire develop team in ISSS,

(c) respondent characteristics from the interview data sets and

(d) interviewer characteristics from a questionnaire administered to all interviewers working on the project.

The distributions of variables that were included in the final multilevel models from each of these sources are listed in Table 1.

The field type variable identifies four different types of response that are existing in Blaise. Single response questions are those where only one response is selected from a list (for example yes–no

questions, gender), as opposed to questions where multiple responses are possible (e.g. 'selected all that applied'), those that allow for a numeric answer (e.g. 'What is your birth year? When did your family move to your current residence?') and questions with open answers. Although these simplified categorizations do not capture all the characteristics of questions, they do reflect different cognitive processes that are necessary to derive answers, and further may result in response time. We expect all three closed question formats would have shorter response time than open questions, and single choice questions to be answered faster than those with multiple-response categories. We also expect items requiring numeric responses to take longer than single response questions.

The field sequence number reflects the actual position of questions answered by the respondent up to that particular field. This indicator varies by respondent, the number of conditional questions and loops that are asked of each respondent.

Since question reading time is an important part of our response time measure by active timer, the word count serves as a key controlling factor. Interviewer instructions, show cards, checks indicate that some actions needed to be taken by the interviewer. Show cards add administration time to show the card to the respondent and also add some reading time. Interviewer instructions may be read in early administrations of the survey but used less as the interviewer becomes more familiar with the instrument. When the inconsistency check functions are active, interviewers need to spend more time to deal with it.

Information on respondent characteristics was added from the CFPS survey dataset. Although the questionnaire contains plenty of data, we restricted ourselves to variables that were previously found to be related to response times. Respondents' age, education was added to our model. We believed that the interviews that were conducted in local dialect may take longer than those conducted in official language(Putonghua) in China, so language used during the interview was added as a control. Since many question in Family financial question is aggregating information on a household level, the former current married groups are likely to have more complex family circumstances. Though marital status was not found stable effect on response time, we still keep it in our model as a control.

We expect interviewers' age and education to play a role in the question administration speed. Similarly, we expect interviewer project experience to be associated with response times, with more experienced interviewers being faster in administering the questionnaire. Interviewer Sequence number is another indicator that reflecting familiarity with the task of administering survey questions. We expect a decline in response time as the number going up. "Complete interviews by an interviewer in the same day" is served as indication of time pressure. Since most of the interviewers live in different community form the survey area, they have motivation to finish more interviews in a travel. And this motivation may result in speeding up in a single interview.

We explore several interactions, e.g. to see whether respondent and interviewer gender different might affect item level times.

Our dependent variable active response time is extracted form Blaise Audit Trail file by AT Report software developed Survey Research Center in University of Michigan. Time measures are recorded in milliseconds in the Blaise audit trails. To increase readability we report time in seconds. We report unadjusted measures of time without controlling for baseline speed for both interviewer and repondent, as is done in many response time studies.

Table 1. Summary of field, respondent and interviewer variables used in the final multilevel models

Variable	Description	% of field
Field haracteristics		
Field squence number	Counter for where the item was asked in each interview	Mean:47.63
Word count	Number of words in field	Mean:18.04
Question type 1	Open	3.42
2	Multiple reponse	8.4
3	Integer	48.25
4	Single reponse	39.93
Flag: showcard	Dummy variable, 1=show card for field	3.31
Flag: F1(QbyQ)	Dummy variable,1=question help available	44.19
Flag: Interviewer instruction	Dummy variable,1=interviewer instruction available for the field	17.49
Flag: check	Dummy variable,1=check available for the field	7.64
Language of interview	Dummy variable, 1=local dialect	25.40
Respondent characteristics		
Age		48.33
Gender	Dummy variable,1=Male	49.80
Marital status	Dummy variable,1=Married	91.26
Education 1	Illiteracy	27.25
2	Primary school graduate	22.90
3	Middle shcool graduate	30.48
4	High shool graduate	16.87
5	Collega graduate and more	2.50
Interviwer characteristics		
Age		Mean:29.02
Gender	Dummy variable,1=Male	55.56
Marital status	Dummy variable,1=Married	50.53
Education		
1	Less than high shool	20.50
2	High shool graduate	27.78
3	Collega graduate and more	42.62
4	Master	9.10
Prior CFPS Experience	Dummy variable,1=yes	60.14
Squrence number by interviewer	Counter for where the interview conducted by each interviewer	47.62
Complete Iws per day	Interviews conducted by the same interviewer in the same day	5.66

4. Methods

Response times are recorded for each question and for each individual respondent conducted by each interviewer in a survey. As a result, response times to survey questions are cross-classified by the interviewer, the respondents and the questions. In other words, the response times from the same interview will not be independent, intra-respondent correlation and/or intra-interviewer can produce a "design effect". Failure to account for such clustering would probably underestimate the standard errors and bias inference that is made from the analyses. Thus, we need to take into account this cross-classified structure when doing analysis on response time.

Given the nature of our data set, unweighted multilevel linear mixed model is used to explore the factors that affect response times. Furthermore, the multilevel model is suitable for our purpose to examine the contribution of the interviewer characteristics on response times.

We first fit the null models, not including any covariates, to ascertain the proportion of variance that is accounted for by each of the levels in the hierarchy. The model can be specified as

$$y_{ijk} = \beta_0 + u_k + u_{jk} + \varepsilon_{ijk}$$

where y_{ijk} is the field time (reported in seconds) for item *i* nested within respondent *j* interviewed by interviewer *k*, $u_k$is the random effect that is associated with the interviewer, u_{jk} is the random effect that is associated with the respondent and$\varepsilon_{ijk}$is the residual variability that is associated with each item *i*, again nested within respondent *j* interviewed by interviewer *k*. All random effects are assumed to follow a normal distribution with$u_{jk} \sim N(0, \sigma_u^2)$ and $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$

Using this, we estimate the intraclass correlation coefficient for interviewer level of the model, as follows:

$$ICC_k = \frac{\sigma_k^2}{\sigma_k^2 + \sigma_{jk}^2 + \sigma_{ijk}^2}$$

and for respondent

$$ICC_{jk} = \frac{\sigma_k^2 + \sigma_{jk}^2}{\sigma_k^2 + \sigma_{jk}^2 + \sigma_{ijk}^2}$$

Where σ_k^2 is the variance of the random effects that are associated with interviewers, $\sigma_{jk}^2$is the variance that is associated with respondents and $\sigma_{ijk}^2$the variance that is associated with items. ICC_k is an estimate of the random variation in response time at the interviewer level, and ICC_{jk} is an estimate of the proportion of random variation at the respondent level.

In order to explore the sources of this variation in more detail, we fit three successive models, first examining field characteristics then adding respondent characteristics and finally adding interviewer characteristics. The full model can be expressed as:

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + \beta_2 V_{jk} + \beta_3 z_k + u_k + u_{jk} + \varepsilon_{ijk}$$

wherex_{ijk} are a vector of covariates at the field level, V_{jk} are respondent level covariates and z_k

are interviewer level covariates.

All models were fitted by using the *xtmixed* procedure in Stata. Since the minimum observations per group in respondent level is 12(see table 2), we used restricted maximum likelihood estimation yielded equivalent results.

Table 2 Grouping result

Group Variable	No. of Groups	Observations per Groups		
		Minimum	Average	Maximum
Interviewer	302	41	875.6	2830
respondent	12963	12	20.4	25

5. Result

5.1 Model fit and estimated random effects

The results of various specifications of random-intercept models without and with covariates are presented in Table 3. From the intraclass correlations, it can be seen that a major portion of the variation in the response times is the residual variance at the individual item level. Respondent and interviewer are accounting for only modest of the variation in field times. Interviewers contribute about 3.1% of the variation and respondents about 4.7%. LR test vs. linear regression is statistically significant $\text{Prob} > \text{chi2} = 0.0000$.

To account for variations across field, we included field characteristics in Model 1. And we found the fixed effects accounts for 4.8% of the variation at the field level (comparing σ^2_{ijk} for model 1 with $\sigma^2_{ijk}$for model 0). Likelihood ratio tests reveal that the addition of these covariates produces statistically significant .p<0:0001.

A set of respondent level fixed effects were added into Model 2. As shown in Table 3, these variables

account for a modest proportion (about 4%) of the variation at the respondent level. However, given the large number of cases, the model fit is significantly .p<0:0001/ improved over model.

Finally we add a set of interviewer level fixed effects to the models. Here the residual reduce about 22%, implying that part of the variation due to interviewers is explained by the interviewer level variables. The likelihood ratio tests for the addition of these fixed effects are significant.

An examination of the coefficients for the fixed effects across the nested models reveals little change in values with the addition of the respondent and interviewer level effects. We thus present estimated coefficients for only the final models, including covariates at all three levels. We describe the effects of each set of variables below.

Table 3 Estimated variance components, intraclass correslation of multilevel model for response time analysis

Model		Result
0,Null model	Variance components	
	σ_{ijk}^{2}	889.26
	σ_{jk}^{2}	15.25
	σ_{k}^{2}	28.60
	Intraclass correlation	
	ICC_k	0.031
	ICC_{jk}	0.047
1,field characteristics	Variance components	
	σ_{ijk}^{2}	846.55
	σ_{jk}^{2}	17.22
	σ_{k}^{2}	28.47
	Intraclass correlation	
	ICC_k	0.032
	ICC_{jk}	0.051
2,field and respondent characteristics	Variance components	
	σ_{ijk}^{2}	846.55
	σ_{jk}^{2}	16.55
	σ_{k}^{2}	28.59
	Intraclass correlation	
	ICC_k	0.032
	ICC_{jk}	0.051
3,field, repondent and interviewer characteristics	Variance components	
	σ_{ijk}^{2}	846.53
	σ_{jk}^{2}	14.35
	σ_{k}^{2}	22.40
	Intraclass correlation	
	ICC_k	0.025
	ICC_{jk}	0.042

Table 4. Estimated final multilevel models including field, respondent and interviewer characteristics (dependent variable: time on field in seconds)

	Coef.	Std. Err.	[95% Conf.	Interval]
Field characteristics				
Field squence number	(0.01)**	0.002	-0.01	0.00
Word count	0.18**	0.008	0.16	0.19
Question type				
Open	—	—	—	—
Multiple reponse	1.38**	0.451	0.50	2.27
Integer	(8.17)**	0.361	-8.88	-7.46
Single reponse	(14.35)**	0.368	-15.07	-13.63
Flag: showcard	1.65**	0.447	0.77	2.53
Flag: F1(QbyQ)	(5.43)**	0.132	-5.68	-5.17
Flag: Interviewer instruction	(7.11)**	0.180	-7.46	-6.75
Flag: check	(7.23)**	0.240	-7.70	-6.76
Respondent characteristics				
Age	0.05**	0.005	0.04	0.06
Gender (1=male)	(0.51)**	0.141	-0.78	-0.23
Marital status(1=married)	0.56*	0.247	0.04	1.01
Education				
Illiteracy	—	—	—	—
Primary school graduate	(0.41)*	0.196	-0.82	-0.05
Middle shcool graduate	(0.47)*	0.197	-0.90	-0.13
High shool graduate	(0.83)**	0.234	-1.38	-0.46
Collega graduate and more	(1.25)**	0.475	-2.28	-0.42
Language	2.67**	0.233	-3.20	-3.67
Interviwer characteristics				
Age	0.06	0.061	-0.06	0.17
Gender (1=male)	(0.17)	0.578	-1.32	0.95
Marital status(1=married)	(0.08)	0.847	-1.72	1.60
Education				
Less than high shool	—	—	—	—
High shool graduate	(2.97)**	0.865	-4.73	-1.34
Collega graduate and more	(2.27)**	0.853	-4.02	-0.68
Master	(2.50)*	1.172	-5.07	-0.48
Prior CFPS Experience (1=yes)	(3.04)**	0.591	-4.21	-1.90
Squrence number by interviewer	(0.07)**	0.004	-0.08	-0.06
Complete Iws per day	(0.35)**	0.018	-0.26	-0.19
_cons	31.82**	1.859	28.81	36.10

* *p*<0.05; ** *p*<0.01

5.2 Impact of field characteristics

Given the large number of observations, it is not surprising that all field level variables reach statistical significance in the model. Here are some main findings from the fixed effects of item level predictors in the models. We can see from Table 4 that the length of the question (measured in the number of words) is positively associated with completion time, but this effect is relatively small, with each additional word adding about a fifth of a second to the time.

We distinguished four question types in our model, as expected open responses (as reference category) take significant longer on average than single response and integer. It is surprising that items with multiple responses take a bit longer time than open responses, even after controlling for the number of words in the field and other factors. In CFPS family questionnaire, those questions with multiple response usually come with a larger number of answer categories might impose greater burden on respondents' working memories. Fields requiring numeric entry take significant longer than single response fields, suggesting that the judgment process may be harder for respondent and administrate work is more complex than fixed choice item for interviewer.

Conditioning on other variables in the model, the field sequence number shows significant negative effect on response time. This finding suggests that interviewers may speed up their delivery and respondents answer more quickly as they got closer to the end of the questionnaire, though the effect of sequence number is modest.

The remaining variables in Table 4 are flags for characteristics of questions that affect the response time. First, hold the other factor constant, items containing show card take slightly longer time on average than those that do not. Fields with show card tend to have more response categories or more complex question. The result shows that it takes the respondent more time to come to an answer.
Second, items containing interviewer instructions take less time on average than those that do not. This is somewhat surprising, as we expected that the presence of instructions indicated a more complex question or the need for further probing by the interviewer. This may also suggest that the interviewers may not be reading the instructions as intended.

Finally, fields with question-by-question help take about 5 second shorter to administer than those without such help. We study the data to find out whether the help screen was actually accessed, out data suggests that the use of such help by interviewers is quite rare, similar with the other study (e.g. Couper et al.,1997).

5.3 Impact of respondent characteristics

The next set of predictors in the models in Table 4 is respondent characteristics. We have already noted that the item level coefficients show little change with the addition of respondent level effects. Our interest here is whether there is systematic variation in administration time by respondent characteristics, controlling for characteristics of the items.

Our analysis shows the expected influence of age and education on response time, which is consistent with earlier findings in the literatures (Couper and Kreuter, 2013). Holding everything else constant, older people and those with less education take longer to answer comparable questions than those who are younger or better educated.

It is not surprising to see that, holding everything else constant, respondents who are currently married significantly taking longer time on single question than those who are not. As we expected, those who were married face more complex recall questions, which in turn would lead to a longer response time.

Finally, there is significant association of language of interview with time. Surveys that are administered in local dialect take longer than those administered in official language, in part because it takes time to translate the question to the respondent.

5.4. Impact of interviewer characteristics

The final set of covariates in Table 4 relates to interviewer characteristics. As noted earlier, interviewers account for a proportion of the overall variation in response times, and adding fixed effects at the interviewer level explain one-fifth of the variation. This suggests that interviewers are having relatively influence on response times. We find that interviewers' education is significantly related to completion time. In part this may reflect the fact that there is more variation in interviewers' education than in respondents' education. Interviewers' age has positive effects on response times, but does not reach significance.

We believed that interviewers who have more experience conducting interview have an advantage in completing our surveys because they are more familiar with instrument. We used sequence number of interview as the indicator of experience. As we expected it is associated with less response time. The findings suggest that the interviewers do gain some experience to administer the instrument as time goes by.

We have an assumption that if there are too many interviews need to be finished in a day, the interviewer will feel some time pressure. Under such pressure, the interviewers tend to speed up the interview to deal with the time constraint. Our result proved this assumption, though the effect is modest, only about one third second per question as adding one more interview.

6. Discussion

We have presented analyses with item level response time as the dependent variable, to explore potential predictors of response times from item, respondent and interviewer level. Our exploratory study has two major findings:

First, we found there are significant random effects and systematic covariation of characteristics at all levels with the response time. The effect of item characteristics can vary across respondents and interviewers. Our work proved that the nested nature of the data should not be ignored. Otherwise it can lead to underestimation of the error terms.

Second, our results parallel some of prior finding on respondent characteristics and item characteristics on response times. And we found CFPS interviewers appear to contribute independently to the response times, and their measured demographic characteristics and experience explain a considerable fraction of the interviewer variance component.

Of course, there are probably many additional variables that affect response times that we were unable to incorporate in our models. The number of covariates on the item level is quite limited, information about the content and nature of the questions were not included in the models. Field features such as the complexity of a question, complexity of response categories, the length of recall periods or the sensitivity of particular items are known to affect response times. Only a limited number of respondent characteristics were available in this model. In addition we did not control the natural rate of speech for interviewer which has proved to be associated with the response time.

Despite the limitations of this study, the replication of known effects with this large field-based data set provides encouragement for us to implement the result during the data collection period. Even though longer response times do not guarantee good data quality or good interviewer performance, shorter response times than average do signal quicker processing in one or more components of the survey response framework and call for attention from survey researchers. We can compare the true values with the predicted values, flagged the sample above or below certain thresholds. If needed, we can arrange other more expensive method to evaluation the performance of the interviewer under a more targeting strategy.

The overall goal is to find ways to use existing data and paradata in an efficient manner to help to evaluate and improve the quality of survey data collection. This work only represents a small step in this direction.

References

Bassili, J. N. and Scott, B. S. (1996) Response latency as a signal to question problems in survey research. *Public Opinion Quarterly,* 60, 390–399.

Couper, M. P., Hansen, S. E. and Sadosky, S. A. (1997) Evaluating interviewer performance in a CAPI survey.

Couper,M. P. (2000).Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64,464–494.

Couper, M. P., Kreuter Frauke(2013), *Journal of the Royal. Statistical Society*, A 176, Part 1, pp. 271–286

Fazio, R. H. (1990) A practical guide to the use of response latency in social psychological research. In *Review of Personality and Social Psychology*, vol. 11, *Research Methods in Personality and Social Research* (eds C.Hendrick and M. S. Clark), pp. 74–97. Newbury Park: Sage.

Mulligan, K., Grant, J. T., Mockabee, S. T., & Monson, J. Q. (2003). Response latency methodology for survey research: Measurement and modeling strategies. *Political Analysis*, 11, 289–301.

Salthouse, T. A. (1991). Theoretical perspectives on cognitive aging. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schwarz, N., Park, D., Kna ̈uper, B., & Sudman, S. (1999). Cognition, aging, and self-reports. Washington, DC: Psychology Press.

Yan, T. and Tourangeau, R. (2008) Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. *Appl. Cogn. Psychol.*, 22, 51–68.