*Journal of Survey Statistics and Methodology (2019) **0**, 1–21*

LANGUAGE PROFICIENCY AMONG RESPONDENTS: IMPLICATIONS FOR DATA QUALITY IN A LONGITUDINAL FACE-TO-FACE SURVEY

ALEXANDER WENZ*
TAREK AL BAGHAL
ALESSANDRA GAIA

When surveying immigrant populations or ethnic minority groups, it is important for survey researchers to consider that respondents might vary in their level of language proficiency. While survey translations might be offered, they are usually available for a limited number of languages, and even then, non-native speakers may not utilize questionnaires translated into their native language. This article examines the impact of language proficiency among respondents interviewed in English on survey data quality. We use data from Understanding Society: The United Kingdom Household Longitudinal Study (UKHLS) to examine five indicators of data quality, including "don't know" responding, primacy effects, straightlining in grids, nonresponse to a self-completion survey component, and change in response across survey waves. Respondents were asked whether they are native speakers of English; non-native speakers were subsequently asked to self-rate whether they have any difficulties speaking or reading English. Results suggest that non-native speakers provide lower data quality for four of the five quality indicators

Alexander Wenz is with the Collaborative Research Center SFB 884 "Political Economy of Reforms," University of Mannheim, B6, 30-32, 68131 Mannheim, Germany and the Institute for Social and Economic Research (ISER), University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK. Tarek Al Baghal is with the Institute for Social and Economic Research (ISER), University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK. Alessandra Gaia is with the Department of Sociology, University of Milan-Bicocca, Via Bicocca degli Arcimboldi 26, 20126 Milano, Italy and the Institute for Social and Economic Research (ISER), University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK. This work was supported by the Economic and Social Research Council [ES/F000537/1] for "Understanding Society: The UK Household Longitudinal Study" (UKHLS). The UKHLS data are available from the UK Data Service at https://discover.ukdataservice.ac.uk/catalogue/? sn=6614, last accessed October 23, 2019.
*Address correspondence to Alexander Wenz, Collaborative Research Center SFB 884 "Political Economy of Reforms," University of Mannheim, B6, 30-32, 68131 Mannheim, Germany; E-mail: a.wenz@uni-mannheim.de.

doi: 10.1093/jssam/smz045

we examined. We find that non-native respondents have higher nonresponse rates to the self-completion section and are more likely to report change across waves, select the primary response option, and show straightlining response behavior in grids. Furthermore, primacy effects and nonresponse rates to the self-completion section vary by self-rated level of language proficiency. No significant effects were found with regard to "don't know" responding between native and non-native speakers.

KEYWORDS: Data quality; Language proficiency; Longitudinal survey; Measurement error.

1. INTRODUCTION

When surveying immigrant populations or ethnic minority groups, it is important to consider that respondents might vary in how well they are able to speak, read, or write in the survey language. Generally, surveys try to include respondents with language problems in the same language as all respondents, although other techniques are sometimes used. For example, respondents not speaking the survey language natively may be provided translated survey instruments, such as in written form for the respondent to complete on their own or through multilingual interviewers or accompanying translators in personal interviews. A problem with translations, however, is that they are only available for a very limited number of languages due to the costs of developing and then administering translated survey instruments. Therefore, not all non-native speakers may be able to take advantage of translated questionnaires and may be interviewed in a language they do not speak well.

When answering survey questions, respondents generally go through four stages of the response process: they need to understand the question, retrieve relevant information, integrate the information to form a judgement, and map their response to the available response options (Tourangeau 1984; Tourangeau, Rips, and Rasinski 2000). Respondents who complete surveys in a non-native language that they have difficulty with might have problems with this process, in particular with understanding and interpreting survey questions. Questions that are complex and put a greater cognitive burden on the respondent, such as those having a complex syntax and containing ambiguous or rarely used words, might be especially difficult to process for non-native speakers of the survey language (Lenzner, Kaczmirek, and Lenzner 2010). The limited question comprehension might also carry over into problems with other stages of the response process: if respondents with language problems do not correctly understand a question, they might also retrieve incorrect information, make incorrect judgments, and map their answer to an incorrect response category.

Previous research using cognitive interviews among non-native respondents found that a large proportion of survey comprehension issues are due to syntax or lexical items (Park, Sha, and Willis 2016), for example questions containing colloquial expressions or terms that are less frequently used (Gray, D'Ardenne, Balarajan, and Uhrig 2011), or due to differences in how concepts are expressed in the different languages. There have been other studies showing that respondents with different cultural or racial/ethnic backgrounds vary in how they understand and interpret survey questions (Warnecke, Johnson, Chávez, Sudman, O'Rourke, et al. 1997; Harkness, van de Vijver, and Mohler 2003; Holbrook, Cho, and Johnson 2006) and how they respond to questions (Johnson, O'Rourke, Chavez, Sudman, Warnecke, et al. 1997; Johnson, Cho, Holbrook, O'Rourke, Warnecke, et al. 2006). At least in some of these instances, this might be due to different language structure. Similarly, there has been evidence that bilingual respondents interpret and respond to survey questions differently depending on the interview language (Peytcheva 2008).

If they are having difficulty carrying out the response process, respondents with language difficulties might be more likely to select inaccurate answers that do not represent their "true" response. Similarly, these respondents may be more likely to select the "don't know" response option due to confusion, not because they do not know an answer; if they had understood, they might have provided a substantive response. An alternative strategy of non-native respondents to cope with comprehension difficulties might be to use other least-effort shortcuts when providing a response, such as always selecting the first or last response option or selecting identical response options in grids.

While quantitative research on language proficiency of respondents is rather limited, a recent study by Kleiner, Lipps, and Ferrez (2015) assessed the response quality of respondents with different language abilities in two telephone surveys in Switzerland. The authors found that foreign-born respondents from countries that did not share one of the Swiss national languages had consistently lower data quality than Swiss respondents, including higher rates of "don't know" responding, straightlining, recency effects and extreme responding, which they attributed to reduced language proficiency and lower motivation of these respondents.

Whether the survey is presented visually or aurally might also play an important role and might interact with individual language proficiency. Differences in speaking, reading, or writing the survey language might affect whether respondents prefer to participate in interviewer- or self-administered surveys. For example, respondents who are more proficient in reading and writing than speaking might prefer to complete self-administered surveys, while those who are more proficient in speaking might prefer to answer survey questions in personal interviews. Among interviewer-administered modes, face-to-face interviews might be preferred over telephone interviews as non-verbal communication from interviewers might also aid the respondent's question comprehension (de Leeuw 1992). For personal interviews, the

interviewers themselves might also be an important factor for data quality (West and Blom 2017). Interviewers with more experience, in particular working with non-native respondents, might be able to elicit higher-quality responses from respondents with limited proficiency in the survey language.

This article examines the impact of language proficiency on survey data quality in a large-scale longitudinal face-to-face survey in the United Kingdom. We analyze the survey data of respondents who completed the interview in English and did not use any translated survey materials. Our article extends the earlier study by Kleiner et al. (2015) in a different mode and cultural setting to further understand the impact of language proficiency on response quality. While Kleiner et al. (2015) used nationality and best-mastered language as proxies for language ability, we employ more direct measures of language proficiency, differentiating between levels of language difficulties among non-native respondents. We also examine additional indicators of data quality that are relevant to a number of studies: nonresponse to a self-completion survey component, which is implemented in many face-to-face surveys for questions that may be better presented visually or are prone to socially desirable responding, and change in response reported over time, an important quality indicator in longitudinal studies.

2. DATA AND METHODS

2.1 Sample

We use data from the first two waves of Understanding Society: The United Kingdom Household Longitudinal Study (UKHLS). It is a large multitopic household survey of the population in the United Kingdom living at residential addresses, with the purpose of collecting high-quality longitudinal data to understand the long-term effects of social and economic change in the country (University of Essex 2019). The UKHLS sample consists of a large general population sample (GPS)—a stratified, clustered sample of households in the UK. The UKHLS uses a probability proportionate to size (PPS) method to select postcode sectors at the first stage with probability relative to the number of households within a sector and then select a set number of addresses within each sector at the second stage, leading to an (approximately) equal probability selection method. Three additional components make up the survey in addition to the main sample: the ethnic minority boost (EMB) sample, the former British Household Panel Survey (BHPS) sample, and the immigrant and ethnic minority boost (IEMB) sample (Knies 2017; Lynn 2009). The EMB sample reflects the largest minority populations in the UK, with at least 1,000 interviews from Indian, Pakistani, Bangladeshi, Caribbean, and African ethnic respondents. The BHPS sample was not integrated until wave two, the IEMB sample was not integrated until wave six of UKHLS, and these are therefore not included in our analysis. The interview is conducted annually among all

eligible household members age sixteen and older. In waves one and two, data were collected using face-to-face computer-assisted personal interviewing (CAPI), with some questions administered through a paper self-completion questionnaire, including questions on health, alcohol consumption, the environment, neighborhood, friendships, and relationships. Most respondents completed the self-completion questionnaire after their face-to-face interview, although some respondents were invited to complete it before their interview while other household members were being interviewed. They were able to return the questionnaire to the interviewer at the end of their visit, have it ready for collection by the interviewer after a few days, or return it by post. While help from other household members was discouraged, interviewers were available to assist if needed. For more details on the survey design and fieldwork, see the study documentation available at www.understandingsociety.ac.uk/documentation/mainstage.

Respondents who were not proficient in English were able to choose whether to complete the interview and the self-completion questionnaire in English or in a translated version. Overall, the use of translated interviews was rather low: for example, in wave one, only 456 translated individual questionnaires were used out of 50,994 interviews (i.e., 0.8 percent). In our analysis, we only use data from interviews conducted in English. Furthermore, we excluded respondents who identify Welsh as their native language from the analysis sample, even if interviewed in English. This is because many Welsh native speakers, although not all, are likely to speak English on a native level, given the ubiquity of English in Wales.

The household response rate for the GPS, with households completing at least one full interview, is 57.3 percent, and the individual response rate in responding households is 81.8 percent at wave one (Knies 2017). The re-interview rate at wave two (i.e., the proportion of wave one respondents who completed a full interview at wave two excluding respondents who became ineligible) is 74.6 percent. The EMB sample had lower response rates: the household response rate is 39.9 percent, and the individual response rate in responding households is 72.4 percent at wave one. The re-interview rate at wave two is 62.2 percent. Data for wave one were collected between January 2009 and March 2011; data for wave two were collected between January 2010 and March 2012.

2.2 Measuring Language Proficiency

All survey respondents at wave one were asked, "Is English your first language?" ("yes," "no") (Q5; the question numbers in parentheses index the questions in appendix A that were used to create the independent and dependent variables; see online supplementary material). Following this question, non-native English speakers were asked, "Do you have any difficulty speaking English to people for day-to-day activities such as shopping or taking the

bus?" ("yes," "no") (Q6) and "Do you have any difficulty reading formal letters or documents written in English?" ("yes," "no") (Q7). There were 6,609 respondents (13.9 percent) who indicated that they are non-native speakers of English, of which 1,235 respondents (18.7 percent of non-natives, 2.6 percent of total) indicated they have difficulty speaking English, and 1,553 respondents (23.5 percent of non-natives; 3.3 percent of total) indicated that they have difficulty reading English.

2.3 Measuring Data Quality

We use five indicators of data quality, some of which were also employed by Kleiner et al. (2015), with alterations to fit the different design aspects of UKHLS. For the first four indicators, only data from the first wave of UKHLS are used; for the fifth indicator, data from the first two waves of UKHLS are used. Two measures that previous research used but are not employed here are those of extreme and mid-five responding, as the number of questions asked in UKHLS using 0–10–point scales was not sufficiently large to replicate these analyses (three questions).

2.3.1 "Don't know" response. "Don't know" responses are examined as in Kleiner et al. (2015): for initial cross-group difference, the proportion of "don't know" responses for each individual are calculated across questions. We then compute a dichotomous measure for each question indicating whether a "don't know" response or some other response has been selected by the respondent. Given the large number of questions in UKHLS and the variation in questions asked to respondents due to routing on prior answers, a subset of question modules asked by the interviewer are used where most questions are asked of most respondents. These include seventy-three questions on receipt of benefits, health and disability, family networks, harassment, environmental behavior, and politics (Q8–Q13; Q26–Q28; Q31–Q33; Q36–Q59; Q61–Q84; Q86–Q87). Most of the questions are factual, with a small number being attitudinal (e.g., political party support).

2.3.2 Primacy effects. Primacy effects are based on responses where the first response option was selected for interviewer-administered questions with four or more response options. UKHLS frequently relies on show cards for such sets of questions, and to ensure comparability, all of the questions employed in the analysis use show cards (twenty questions: Q17–Q20, Q22–Q25, Q86, Q105). The visual presentation of show cards suggests primacy effects (Krosnick and Alwin 1987; Schwarz, Strack, Hippler, and Bishop 1991) and contrasts with the aural presentation in Kleiner et al. (2015), which identified recency effects. Although the analysis of primacy effects is ideally carried out on questions with nominal scales containing a large number of response

options, all questions in this analysis use ordinal scales because we were constrained by the types of questions employed in UKHLS. For initial cross-group differences, the proportion of primacy responses for each individual are calculated across questions, as before. We also create a dichotomous variable for each of the twenty questions indicating whether the respondent gave the first response or not.

2.3.3 Straightlining. Straightlining, in which the respondent gives the same response for each item in a grid, is also examined as in Kleiner et al. (2015), with the important difference being that the indicators come from visually designed grids. Straightlining in grids has been identified in particular as a potential source of reduced data quality in visually designed modes (Couper, Tourangeau, Conrad, and Zhang 2013) and is indicated in UKHLS by two sets of questions presented in grids as part of the self-completion component. One grid contained eight questions regarding attitudes towards the respondent's neighborhood (Q118), and the other seven questions were about personal well-being (Q120). Straightlining is identified for each grid (i.e., a respondent gives the same response for each item within a single grid). Two dichotomous indicators of straightlining are then computed, whether a respondent did or did not straightline in each grid, so that a proportion of straightlining can be included for initial cross-group differences and individual outcomes used for later analyses. The questions asked by the interviewer that could be used to indicate straightlining are also those largely used to indicate primacy. Including these questions in both primacy and straightlining indicators would suggest two separate behaviors are taking place, when only one is taking place, and only one outcome is being affected.

Two other design aspects of UKHLS allow for data quality measures not possible in Kleiner et al. (2015), which may be of particular importance for a number of studies.

2.3.4 Nonresponse to self-completion section. One quality indicator is whether the respondent completed the self-completion section of the survey or declined to do so. Many face-to-face surveys include self-completion components to ask questions that may be better presented visually or are prone to socially desirable responding. These components require a different set of language skills, particularly those involved with reading and comprehension, since interviewer assistance is more limited in this mode. If respondents with language difficulties are less likely to respond to the self-completion section, then a large number of the same questions may be missing for this subset of the sample.

2.3.5 Change in response. Change in response is an important indicator of data quality in longitudinal studies, with a significant amount of indicated

change being spurious over-reports (Jäckle 2009; Al Baghal 2017). If change is indicated at differential rates across language proficiency groups, then the longitudinal data quality may also vary. A total of fifty-six questions are used to explore change across the first two waves of UKHLS, with a dichotomous outcome that indicates whether any change of substantive responses occurred across the two waves or not (Q14–Q26, Q29–Q30, Q34–Q35, Q60, Q61, Q70–Q72, Q88–Q96, Q98–Q117, Q119). Only questions from the interviewer-administered component are considered, given the selection differences in self-completion response. The fifty-six questions used are selected on the basis of being asked of most respondents and being categorical in nature. The latter criterion is set because change in interval types of responses is more difficult to define (Lugtig and Lensvelt-Mulders 2014; Al Baghal 2017). A number of the included questions are attitudinal (e.g., job and life satisfaction, political party support), in addition to factual questions on health, employment, caring for others, and benefit receipt.

2.4 Question Coding

For the analyses of "don't know" response and change in response across waves, indicated at the question level, we are able to control for question characteristics that have been coded by NatCen Social Research (D'Ardenne, Collins, Gray, Jessop, and Pilley 2017). First, we include three measures of question complexity: whether the question stem includes lengthy instructions, introductions, or explanations (complex question stem); whether the questionnaire includes additional explanations beyond those included in the question stem (extra information); and whether the question involves any mental calculation (computation). Second, we include three measures of the response format: whether the question is part of a question battery that all use the same response scale (battery of scalars); whether the question includes a rating scale (rating scale); and whether it has five or more response options (5+ response options). Finally, we include measures of whether the question asks about illegal or illicit behaviors (fear of disclosure), whether the question has some response options that are more socially desirable than others (social desirability), and whether the question includes a show card (show card). The questions were coded by two coders independently; in the case of disagreement, a final code was allocated by a third coder. The intercoder reliability was high for all codes (88.9 percent and higher) with one exception (social desirability: 59.6 percent). We also control for whether the question is attitudinal or factual.

2.5 Analysis Methods

The agreement to do the self-completion section occurs only once in the interview (at the respondent level), and so it is only clustered within interviewers. The remaining quality measures are indicated in models at the item level.

"Don't know" response, primacy, straightlining, and change in response are analyzed both at the respondent level, where outcomes are only nested within interviewers, and at the question level, where outcomes are nested within both respondents and interviewers, necessitating a three-level model. As the questions are asked both across respondents and interviewers, a cross-classified multilevel model is used for all models utilizing item-level data (e.g., Yan and Tourangeau 2008).

For "don't know" and change models, examining the measures at the question level allows for including question characteristics in the model, described previously, to further disentangle possible language effects. However, these question characteristics are not useful for the primacy and straightlining models, as the variables used in these analyses come from a question battery on the same topic and with the same format and, hence, have near exact question coding (e.g., all attitudinal). To maintain comparability with "don't know" and change models, fixed effects for each item are not included in these other models. The model equations and analytic sample for each of the data quality measures are outlined in appendix B of the online supplementary material.

The question for English as a first language is asked of all respondents, where difficulties in speaking and reading English is asked only of those who indicated English is not their first language. Due to this structure, native English speakers who have difficulties speaking or reading cannot be separated out to explore how these difficulties impact data quality across language nativity. Therefore, models exploring English as a first language predicting data quality include all respondents, while models for speaking and reading difficulties include only those having English as a second language. This modeling strategy allows identification of whether differences occur between native and non-native speakers or among non-native speakers with different levels of speaking and/or reading skills.

The multilevel models estimating the impact of language proficiency on change indicators use the same modeling strategy, but the analysis sample is reduced by two factors. First, due to attrition between the first and second wave of UKHLS, some respondents are not available for analysis. Second, to identify interviewer effects, only respondents surveyed by the same interviewer in both the first and second wave are included in the analysis sample. This reduction leads to 24,153 respondents available for analysis of change compared with 50,538 for analysis of the other indicators that use the first wave only. Given the multiple responses per respondent, however, there are still 799,360 observations for analysis of change.

In addition to the indicators of language proficiency, several other variables are included in the multivariate models to control and understand respondent and interviewer effects on data quality. Respondent indicators of sex (Q1), being born in the UK or not (Q3), education (Q4), and age (Q2) are included in all models. For educational attainment, those with less than a professional degree are in the baseline educational category compared with those with a

professional or university degree. A proxy measure for respondent effort comes from the subjective rating by the interviewer of the respondent's cooperation on a five-point scale (Q121). Given the heavily skewed nature of the data towards very good cooperation, data were recoded dichotomously to cooperation being very good or not. For models estimating change, respondents rated as having very good cooperation at both waves are compared with all other respondents (i.e., rated cooperative only in wave two but not in wave one, only in wave one but not in wave two, or in neither wave). Due to an error in paradata capture, survey completion times were not recorded. However, interviewers estimated the interview length at the end of the survey, and this measure is used to indicate time to complete the survey (Q125). Interview length is a possible indicator of respondent-interviewer rapport (e.g., Jenkins, Cappellari, Lynn, Jäckle, and Sala 2006) or could also indicate respondent difficulty in answering the questionnaire. For the model estimating change, the difference in estimated interview length between waves (wave one–wave two) is included to analyze the impact that differential timing in surveys has on change.

Interviewers are not interpenetrated across primary sampling unit (PSU) (i.e., one interviewer represents one PSU). Inclusion of random effects for the interviewer captures the clustering of PSU. To account for possible regional impacts, however, models include the eleven UK government office regions as controls (with North East region used as a baseline) (Q126). Because these are included as controls and not of substantive interest, for presentation purposes, these are not included in the tables. Beyond PSU, stratification is not included in the models as the strata used for the general sample is different but overlapping with EMB sample design; further, including stratification in expectation reduces variance estimates. Hence, the estimates are likely to be more conservative regarding statistical significance.

At the interviewer level, 888 interviewers completed interviews at wave one, with 648 interviewing the same respondent at least once across the first two waves. The interviewer demographics available from the fieldwork agency include age (Q123), sex (Q122), and ethnicity. However, a large number of interviewers refused to disclose their ethnicity (21.8 percent), so interviewer ethnicity will not be considered further. Experience as an interviewer at the fieldwork agency is also included (Q124). Sex, age, and experience at the fieldwork agency are missing for seventeen interviewers, leaving 871 interviewers for analysis.

Table 1 shows the variables described previously and used in the full models to predict data quality for respondents and interviewers by native language status.

There are only three statistically significant differences across native and non-native English speakers. The first, as might be expected, is whether the respondent was born in the UK or not. Data show that 92.4 percent of native English speakers were born in the UK compared with only 8.0 percent of non-

Table 1. Mean/Proportion of Respondent and Interviewer Characteristics by Native Language Status

	Native English (SD; n)	Non-native English (SD; n)
Respondent characteristics		
Female	0.562	0.536
	(0.50; 41,061)	(0.50; 5,787)
UK born	0.924	0.080
	(0.27; 41,056)	(0.31; 5,784)
Age	47.06	37.71
	(18.36; 41,061)	(13.88; 5,787)
University degree	0.204	0.335
	(0.40; 40,672)	(0.47; 5,777)
Professional degree	0.115	0.102
	(0.32; 40,672)	(0.30; 5,777)
Very cooperative	0.775	0.604
	(0.42; 40,960)	(0.49; 5,745)
Interviewer characteristics		
Interviewer-age	57.14	56.60
	(10.17; 870)	(10.65; 615)
Interviewer-female	0.514	0.501
	(0.50; 870)	(0.50; 615)
Years as interviewer	5.73	5.55
	(5.12; 870)	(4.90; 615)
Interview length in minutes	42.75	44.50
	(16.76; 40,278)	(19.19; 5,661)

native English speakers. The other two significant differences have possible relationships with data quality. A significantly smaller proportion of non-native English speakers were rated by the interviews as being very cooperative during the interview (60.4 percent) compared with English native speakers (77.5 percent). Conversely, significantly more non-native speakers have a university degree (33.5 percent) compared with native English speakers (20.4 percent), which potentially has a countervailing impact on data quality.

3. RESULTS

Table 2 shows the percentage of each of the data quality indicators by native language status and language proficiency, and tests for differences in behaviors between categorizations. There are no identified differences in the aggregated indicator of "don't know" response between native and non-native English speakers. There are also no significant differences identified within non-native English speakers, comparing those indicating difficulty speaking or reading to those saying they have no such problems. Although there are directionally

Table 2. Percentage of Respondent Behaviors for Data Quality Indicators

	Native language status		Speaking difficulty (Non-native English)		Reading difficulty (non-native English)	
	Native English	Non-native English	No difficulty	Difficulty	No difficulty	Difficulty
Percent "don't know" response	0.98	1.14	1.08	1.46	1.01	1.62
	$F(1, 46{,}846) = 1.19$ $p = 0.276$		$F(1, 5{,}783) = 0.94$ $p = 0.332$		$F(1, 5{,}785) = 3.00$ $p = 0.083$	
Percent primacy response	16.35	17.03	16.90	17.80	16.82	17.85
	$F(1, 38{,}505) = 38.57$ $p < 0.0001$		$F(1, 5{,}769) = 5.30$ $p = 0.021$		$F(1, 5{,}771) = 10.43$ $p = 0.001$	
Percent straightlining	2.95	4.13	4.04	4.82	3.98	4.95
	$F(1, 38{,}505) = 15.96$ $p < 0.0001$		$F(1, 3{,}801) = 0.59$ $p = 0.443$		$F(1, 3{,}803) = 1.18$ $p = 0.277$	
Percent self-completion nonresponse	12.48	30.07	27.43	45.23	26.33	44.86
	$F(1, 46{,}846) = 1165.58$ $p < 0.0001$		$F(1, 5{,}783) = 106.91$ $p < 0.0001$		$F(1, 5{,}785 = 147.93$ $p < 0.0001$	
Percent response change W1-W2	33.15	35.21	35.44	33.86	35.12	35.56
	$F(1, 22{,}200) = 3.15$ $p = 0.076$		$F(1, 1{,}799) = 0.25$ $p = 0.617$		$F(1, 1{,}800) = 0.02$ $p = 0.875$	

more "don't know" responses for those with lower language proficiency or those who are non-native English speakers. Similarly, there are no significant differences across all comparisons for the aggregated indicator for change in response between the first two waves.

For the other three data quality indicators, however, there is at least one significant difference identified between categorizations of language proficiency. In all such cases, those self-reporting less language proficiency or being non-native speakers display greater percentages of behaviors suggesting lower data quality. Two of those were also shown to be affected by language proficiency in Kleiner et al. (2015). In particular, significantly more primacy is observed among non-native English speakers. Among those non-native speakers, more primacy is observed among those self-rated as having difficulty speaking and difficulty reading compared with those saying they have no difficulty with these (all at least at $p < 0.05$). More straightlining is also found among non-native English speakers compared with native speakers ($p < 0.0001$). However, within non-native speakers, no significant differences are found in straightlining behavior between those expressing difficulty with speaking or reading English compared with those saying they do not have problems with each of these. Although significant, the effect size of the differences in primacy effects and straightlining are small, with Cohen's d between 0.039 and 0.106.

Perhaps the most striking differences in these data quality indicators are the nonresponse rates to the self-completion section of the interview. Nearly 18 percentage points more of non-native English speakers did not respond to the self-completion section than native English speakers (30.1 percent versus 12.5 percent), resulting in a much higher effect size ($d = 0.479$). Within non-native speakers, those saying they have no difficulty with speaking and reading still have nonresponse rates (27.4 percent and 26.3 percent, respectively) much higher than native English speakers. Yet, nonresponse rates among respondents with speaking and reading difficulties, both being close to 45 percent, are significantly higher ($p < 0.0001$) than the rates of non-native English speakers with no language difficulties ($d = 0.382$ for speaking difficulty, $d = 0.398$ for reading difficulty). Such high nonresponse rates are perhaps the clearest evidence of lower data quality among non-native speakers and among respondents with difficulties speaking and reading the survey language especially.

This analysis indicates the possible differences in data quality across native language status and language proficiency. However, these results do not control for a variety of respondent and interview characteristics that may otherwise help explain the observed differences or the lack thereof. Therefore, for each of the five indicators of data quality, multilevel logistic regression models were estimated, including the respondent- and interview-related measures in table 1 as independent variables. The measures of "don't know" response, primacy, straightlining, and change in response are indicated at the question level, so they are nested within respondents and interviewers; the response to the self-

Table 3. Odds Ratios from Multilevel Models Predicting Data Quality Indicators

	"Don't know" response	Primacy	Straightlining	Self-completion nonresponse	Response change W1-W2
Respondent characteristics					
Non-native English	1.041	1.048**	1.181**	1.713**	1.166**
Female	1.232**	1.044**	1.226**	0.959	1.086**
UK born	0.963	0.912**	1.031	0.620**	0.951**
Age	1.014**	1.005**	1.006**	1.007**	0.998**
Education (Baseline: less than professional)					
College degree	0.969	0.809**	0.926*	0.773**	0.839**
Professional	1.028	0.886**	0.984	0.808**	0.898**
Interview characteristics					
Interview length	1.004**	1.001*	0.997**	0.998	1.001*
Respondent cooperative	0.680**	1.005	0.976	0.361**	0.868**
Interviewer-female	1.310**	1.068**	1.016	0.685**	1.014
Interviewer-age	0.989**	1.001	1.000	0.961**	0.998**
Years as interviewer	1.017**	0.998	0.999	1.054**	0.998
Question fixed effects					
Fear of disclosure	1.422**	—	—	—	—
Social desirability	0.489**	—	—	—	0.818**
Rating scale	0.156**	—	—	—	1.705**
Complex question stem	1.603**	—	—	—	1.505**
Extra information	0.950	—	—	—	0.725**
Computation	11.920**	—	—	—	—
Battery of scalars	0.669**	—	—	—	1.313**
5+ response options	1.893**	—	—	—	1.522**
Attitude question	8.237**	—	—	—	3.381**
Show card	0.557**	—	—	—	1.262**
Respondent variance	0.871	0.119	1.177	—	0.324
Interviewer variance	0.469	0.053	0.052	0.386	0.006
n Responses	1,503,331	835,873	75,663	—	751,937
n Respondents	45,732	45,384	38,241	45,385	20,939
n Interviewers	871	871	853	871	640

NOTE.— Exponentiated coefficients. *$p < 0.05$, **$p < 0.01$. The models also control for UK government office region as fixed effects (11-category variable, North East region baseline), but the coefficients are not reported here.

completion component is a single respondent outcome and is nested within interviewers only. Results of these models are presented in table 3.

Of most immediate importance, being a non-native English speaker is a statistically significant predictor for four of the five data quality indicators after controlling for a variety of respondent and survey characteristics. Non-native English speakers are predicted to be more likely to display primacy response selection, straightline in grids, make a change in response across waves, and less likely to respond to the self-completion section. Perhaps not surprisingly, given the findings in table 2, the odds of language impacting data quality are largest in the nonresponse to the self-completion component of the survey. Only the odds of giving a "don't know" response are not significantly impacted by whether respondents are native English speakers, although the estimate is in the same direction as in the other models. The lack of findings of native language status on "don't know" responses is the opposite of that found in Kleiner et al. (2015), who found that non-native speakers generally give more "don't know" answers.

Respondents born in the UK are less likely to display primacy effects, more likely to respond to the self-completion section, and less likely to report a change across waves. Being born in the UK likely has the impact of improving English language even among respondents self-identified as having a different first language. Indeed, only 2.9 percent of UK-born non-native English speakers indicated difficulties speaking, while 20.6 percent of non-native speakers born outside of the UK indicated speaking difficulties ($\chi_1^2 = 182.92$, $p < 0.0001$). Similarly, 3.3 percent of UK-born non-native speakers said they had difficulty reading English compared with 26.0 percent of non-native speakers born elsewhere ($\chi_1^2 = 132.42$, $p < 0.0001$).

Age and education, frequently used as proxies for cognitive ability, are significant and generally in the direction found in previous data quality studies (Knäuper 1999; Schwarz, Park, Knäuper, and Sudman 1999; Kaminska, McCutcheon, and Billiet 2010; Al Baghal 2017). That is, older and less educated respondents are more likely to provide lower data quality responses. These respondents are significantly more likely to provide a "don't know" response, be affected by primacy, straightline, and decline the self-completion section, even after controlling for other factors including language proficiency. While less educated respondents are more likely to report a change across waves, older respondents are less likely to do so. The change in direction of this estimate may be due to the possibility that older respondents genuinely have less change in their lives than younger respondents (Al Baghal 2017).

Another important finding is that respondents who were rated cooperative by interviewers generally have significantly lower odds of providing indicators of lower data quality, mirroring previous results on the impact of respondent motivation (e.g., Kleiner et al. 2015). Cooperativeness has no significant relationship with primacy measures, but all other models estimate higher data quality where respondents are identified as cooperative. Conversely, longer

Table 4. Odds Ratios for Difficulty Indicators Predicting Data Quality Indicators

	"Don't know" response	Primacy	Straightlining	Self-completion nonresponse	Response change W1-W2
Non-native speakers					
Speaking difficulty	1.054	1.088**	0.887	1.847**	1.040
Reading difficulty	1.223*	1.107**	0.969	2.186**	1.089

NOTE.— Exponentiated coefficients. *$p < 0.05$, **$p < 0.01$.

interviews are significantly related to lower data quality in three models ("don't know" response, primacy, and change) but are unrelated or related to higher data quality in the other two. The relationship of longer interviews after controlling for cooperativeness, potentially also related to rapport, may be indicative of greater difficulty experienced, although this needs further exploration.

Question characteristics also show significant relationships with data quality in the "don't know" and change models. However, outcomes are not always consistent across the two models, making some conclusions less clear. Questions that potentially induce fear of disclosure, questions with a complex question stem, questions with five or more response options, and questions requiring computation (adding complexity) have higher odds of producing "don't know" responses than questions without those characteristics. Disclosure risk and complexity are expected causes of increases in item-nonresponse (Tourangeau et al. 2000; Biemer and Lyberg 2003). Of these three characteristics, only items with complex stems were also included in the analytic sample for items exploring change across waves. As with "don't know" responses, questions with complex stems display higher odds of change indicated across waves, suggesting the importance of question complexity on data quality. Questions with extra information provided have significantly lower odds of both "don't know" responses and change across waves. These lower odds are consistent with the complexity interpretation, as these are intended to reduce question complexity. Conversely, attitude questions have significantly higher odds of "don't know" responses and change across waves. There may actually be a greater lack of knowledge about attitude objects, and there may also be greater levels of change in attitudes relative to behaviors. The higher odds for "don't know" responses may also be due to these questions requiring greater cognitive burden in constructing a response (Tourangeau et al. 2000). Along with this burden, some respondents with less strong attitudes will rely on temporarily accessible information (Sudman, Bradburn, and Schwarz 1996), and this information may change across waves, either through different survey characteristics or other respondent characteristics affecting rates of change (Al Baghal 2017).

Questions that may induce socially desirable reporting also have lower odds of "don't know" responses and change across waves, which is somewhat contrary to expectation. Rather than leading to indication of lower data quality, fewer "don't know" responses and less change is found when these questions are asked, controlling for other question factors. One explanation is that the socially desirable responses are apparent and the "correct" choice to select, both within (reducing "don't knows") and across waves (reducing change). Questions with rating scales, that are part of a battery of scalars, or that have a show card are also significantly less likely to have "don't know" responses; however, these types of questions are related to higher reports of change, so the overall indication on data quality is less clear.

Additionally, we were able to further explore if self-rated speaking and reading difficulties among non-native speakers differentiated these respondents on quality indicators. The same models as in table 3 were fitted on the subset of non-native speakers of English, substituting the speaking and reading difficulty indicators for the native language indicator into separate models for each. The same independent variables and modeling structure were used. The results for the independent variables other than the language proficiency indicators suggested a similar pattern as found in table 3. To summarize these additional ten models to the key outcome, the odds ratios estimated for both speaking and reading difficulty measures for each of the data quality indicators are presented in table 4.

Beyond the direct impact of native language status on data quality, those with reading difficulties (but not speaking difficulties) are significantly more likely to give a "don't know" response. Further, self-rated difficulty with speaking and reading English has a statistically significant impact only on primacy and nonresponse self-completion indicators. For these two, not only does native language status impact data quality but also the level of difficulty of non-native speakers has an additional impact on data quality relative to non-native speakers without these difficulties. These results importantly suggest that there is some differentiation by language proficiency. In particular, this is most evident in the instances when reading is required. Primacy effects are expected instead of recency effects due to the reliance on show cards for these questions, and the self-completion section is introduced as a reading exercise. Straightlining, the other indicator using reading, is not significantly different among self-rated ability within non-native speakers. This lack of difference is likely to be, at least partly, due to those respondents having the most difficulty choosing not to do the self-completion section.

4. DISCUSSION

In this article, we examine how language proficiency among respondents affects survey data quality, using data from a large-scale longitudinal

household survey in the United Kingdom. Overall, we find that native English speakers provide better data quality both within and across waves. Being a non-native English speaker has a significant negative effect on four of the five data quality indicators that we examined, when controlling for a variety of respondent and survey characteristics.

The most striking finding is that non-native speakers are much less likely than native English speakers to respond to the self-completion section of the survey, requiring reading and writing skills. The refusal rates also differ by level of language proficiency: among non-native speakers, those indicating difficulties with speaking and reading English are even more likely to refuse than non-native speakers without these problems. It is reasonable to suggest that if those non-native speakers with more difficulties with the survey language are encouraged to complete the self-completion section, in some other ways might data quality suffer, such as an increase in straightlining. Nonresponse to the self-completion section entails a large number of the same set of questions not answered by a large percentage of particular respondents. If this missingness was included in an item-nonresponse analysis, that is comparing the item-nonresponse rates by language proficiency, it is likely that we would have found large differences. Further, questions included in self-completion sections are frequently done so due to the nature of the questions being asked, for example being more sensitive than others. Thus, the large number of questions missing may have particular importance to certain research questions and are missing at much higher rates among a population that may be different in key ways on these measures.

Our results also suggest significant effects of language proficiency on three other data quality indicators. First, we find that non-native speakers are more likely to report change across survey waves than native speakers. Generally, a large amount of change indicated in longitudinal surveys are spurious over-reports (Jäckle 2009). Some of the higher level of change for non-native speakers, however, is likely to be real as they may experience more change when settling in a new country. Still, taking this commonly used indicator of longitudinal data quality along with the others presented suggests the impact of language nativity on outcomes. Interestingly, we do not find significant differences in the amount of change reported between non-native speakers with self-rated language difficulties and those without, which suggests that only non-native language status and not the level of language difficulties contributes to lower data quality. Second, our results suggest that non-native speakers are more likely to select primary response options for questions presented on show cards; among non-native speakers, primacy effects are more likely to occur for those reporting language difficulties with speaking and reading. This finding is in line with Kleiner et al. (2015), who found higher recency effects for respondents with lower language proficiency in telephone surveys. Finally, we find that non-native speakers are significantly more likely to show straightlining response behavior in grids than native speakers; among non-

native speakers, we do not find differences by speaking or reading difficulty. Again, this indicates that non-native speakers contribute to lower data quality, independent of their level of language proficiency.

Surprisingly, our study does not find evidence for higher rates of "don't know" responses among non-native respondents or those with language difficulties, which is contrary to the results of Kleiner et al. (2015), who found that foreign-born respondents provide more "don't know" responses. The inconsistent finding might be explained by the different mode of data collection of the two studies and the associated social conventions. In face-to-face surveys, the locus of control during the interview is usually shared between the respondent and the interviewer, determining the pace and flow of communication (de Leeuw 1992). Telephone interviewers, however, tend to have more control over the interview and try to avoid long silences during the conversation. We might expect that respondents with limited language proficiency are given more time in face-to-face interviews to understand the question and provide a substantive response, whereas they are more likely to be rushed in telephone interviews, resulting in higher rates of non-substantive responses. Further research, however, is needed to replicate these findings and better understand the impact of language proficiency on "don't know" responding.

A limitation of our study is that respondents with language difficulties were able to self-select whether to complete the interview in English or use one of the survey translations, if available. Those who chose to do the survey in English are likely to have different characteristics than those who chose to complete the survey in a translated version. In particular, respondents with the greatest language problems are likely to have chosen the translations, so we might even expect lower data quality among non-native speakers if these respondents were also interviewed in English. Second, we used self-reported measures of language proficiency in our analysis, which might be affected by measurement error, as respondents are likely to overestimate their language abilities. Future studies on language proficiency among survey respondents could use more accurate measures, for example based on a more standardized language test. Finally, our analysis is mainly based on indirect measures of data quality, such as primacy response selection or straightlining in grids. While these indicators provide first evidence for lower data quality among non-native respondents, future research could attempt to capture measurement error more directly, for example by comparing survey data with validation data from administrative records.

Supplementary Materials

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

REFERENCES

Al Baghal, T. (2017), "Last Year Your Answer Was . . .: The Impact of Dependent Interviewing Wording and Survey Factors on Reporting of Change," *Field Methods*, 29, 61–78.

Biemer, P., and L. Lyberg (2003), *Introduction to Survey Quality*, Hoboken, NJ: John Wiley.

Couper, M. P., R. Tourangeau, F. G. Conrad, and C. Zhang (2013), "The Design of Grids in Web Surveys," *Social Science Computer Review*, 31, 322–345.

D'Ardenne, J., D. Collins, M. Gray, C. Jessop, and S. Pilley (2017), "Assessing the Risk of Mode Effects: Review of Proposed Survey Questions for Waves 7-10 of Understanding Society." Understanding Society Working Paper, 2017-04.

Gray, M., J. D'Ardenne, M. Balarajan, and N. Uhrig (2011), "Cognitive Testing of Wave 3 Understanding Society Questions," Understanding Society Working Paper, 2011–03.

Harkness, J. A., F. J. R. van de Vijver, and P. P. Mohler (2003), *Cross-Cultural Survey Methods*, Hoboken, NJ: John Wiley.

Holbrook, A. L., Y. I. Cho, and T. P. Johnson (2006), "The Impact of Question and Respondent Characteristics and Mapping Difficulties," *Public Opinion Quarterly*, 70, 565–595.

Jäckle, A. (2009), "Dependent Interviewing: A Framework and Application to Current Research," in *Methodology of Longitudinal Surveys*, ed. P. Lynn, pp. 93–112, Chichester: John Wiley,

Jenkins, S. P., L. Cappellari, P. Lynn, A. Jäckle, and E. Sala (2006), "Patterns of Consent: Evidence from a General Household Survey," *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 169, 701–722.

Johnson, T. P., Y. I. Cho, A. L. Holbrook, D. O'Rourke, R. B. Warnecke, and N. Chavez (2006), "Cultural Variability in the Effects of Question Design Features on Respondent Comprehension of Health Surveys," *Annals of Epidemiology*, 16, 661–668.

Johnson, T. P., D. P. O'Rourke, N. Chavez, S. Sudman, R. B. Warnecke, and L. Lacey (1997), "Social Cognition and Responses to Survey Questions among Culturally Diverse Populations," in *Survey Measurement and Process Quality*, eds. L. Lyberg, P. Biemer, M. Collins, E. D. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, pp. 87–113, New York: John Wiley.

Kaminska, O., A. L. McCutcheon, and J. Billiet (2010), "Satisficing among Reluctant Respondents in a Cross-National Context," *Public Opinion Quarterly*, 74, 956–984.

Kleiner, B., O. Lipps, and E. Ferrez (2015), "Language Ability and Motivation among Foreigners in Survey Responding," *Journal of Survey Statistics and Methodology*, 3, 339–360.

Knäuper, B. (1999), "The Impact of Age and Education on Response Order Effects in Attitude Measurement," *Public Opinion Quarterly*, 63, 347–370.

Knies, G. (2017), *Understanding Society the UK Household Longitudinal Study, Waves 1-7, User Guide*, Colchester: University of Essex.

Krosnick, J. A., and D. F. Alwin (1987), "An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement," *Public Opinion Quarterly*, 51, 201–219.

de Leeuw, E. D. (1992), *Data Quality in Mail, Telephone, and Face-to-Face Surveys*, Amsterdam: TT-Publicaties.

Lenzner, T., L. Kaczmirek, and A. Lenzner (2010), "Cognitive Burden of Survey Questions and Response Times: A Psycholinguistic Experiment," *Applied Cognitive Psychology*, 24, 1003–1020.

Lugtig, P., and G. J. L. M. Lensvelt-Mulders (2014), "Evaluating the Effect of Dependent Interviewing on the Quality of Measures of Change," *Field Methods*, 26, 172–190.

Lynn, P. (2009), "Sample Design for Understanding Society." *Understanding Society Working Paper*, 2009-01.

Park, H., M. M. Sha, and G. Willis (2016), "Influence of English-Language Proficiency on the Cognitive Processing of Survey Questions," *Field Methods*, 28, 415–430.

Peytcheva, E. (2008), *Language of Administration as a Source of Measurement Error: Implications for Surveys of Immigrants and Cross-Cultural Survey Research, Doctoral Dissertation*, Ann Arbor: University of Michigan.

Schwarz, N., D. Park, B. Knäuper, and S. Sudman (1999), *Aging, Cognition, and Self-Reports*, Washington, DC: Psychology Press.

Schwarz, N., F. Strack, H.-J. Hippler, and G. Bishop (1991), "The Impact of Administration Mode on Response Effects in Survey Measurement," *Applied Cognitive Psychology*, 5, 193–212.

Sudman, S., N. M. Bradburn, and N. Schwarz (1996), *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*, San Francisco: Jossey-Bass.

Tourangeau, R. (1984), "Cognitive Sciences and Survey Methods," in *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*, eds. T. B. Jabine, M. L. Straf, J. M. Tanur, and R. Tourangeau, pp. 73–100, Washington, DC: National Academy Press.

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

University of Essex, Institute for Social and Economic Research. (2019), *Understanding Society: Waves 1-8, 2009-2017 and Harmonised BHPS: Waves 1-18, 1991-2009.* [data collection]. *11th Edition*. UK Data Service. Available at https://doi.org/10.5255/UKDA-SN-6614-12 (last accessed Octobere 23, 2019).

Warnecke, R. B., T. P. Johnson, N. Chávez, S. Sudman, D. P. O'Rourke, L. Lacey, and J. Horm (1997), "Improving Question Wording in Surveys of Culturally Diverse Populations," *Annual Epidemiological Psychology*, 7, 334–342.

West, B. T., and A. G. Blom (2017), "Explaining Interviewer Effects: A Research Synthesis," *Journal of Survey Statistics and Methodology*, 5, 175–211.

Yan, T., and R. Tourangeau (2008), "Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times," *Applied Cognitive Psychology*, 22, 51–68.