Practice of Epidemiology

Interviewer Error Within the Face-to-Face Food Frequency Questionnaire in Large Multisite Epidemiologic Studies

Chengyuan Sun, Bing Guo, Xiang Liu, Xiong Xiao*, and Xing Zhao* on behalf of the China Multi-Ethnic Cohort (CMEC) Collaborative Group

* Correspondence to Dr. Xing Zhao, Department of Epidemiology and Biostatistics, West China School of Public Health and West China Fourth Hospital, Sichuan University, No. 16, Section 3, Renmin South Road, Chengdu, Sichuan, CN 610041 (e-mail: xingzhao@scu.edu.cn); or Dr. Xiong Xiao, Department of Epidemiology and Biostatistics, West China School of Public Health and West China Fourth Hospital, Sichuan University, No. 16, Section 3, Renmin South Road, Chengdu, Sichuan, CN 610041 (e-mail: xiongxiao@scu.edu.cn).

Interviewer error has long been recognized in face-to-face surveys, but little is known about interviewer error within face-to-face food frequency questionnaires, particularly in large multisite epidemiologic studies. Using dietary data from the China Multi-Ethnic Cohort (2018–2019), in which all field interviews were audio recorded, we identified a potentially error-prone sample by outlier detection and further examined the interviewer errors by reviewing these error-prone interviews. Among 174,012 questions for 5,025 error-prone interviews, 13,855 (7.96%) questions were identified with interviewer error, which mainly came from falsification (37.53%), coding error (31.71%), and reading deviation (30.76%). We found that 98.29% of interviewers and 73.71% of respondents had at least 1 error, and half of the errors could be attributed to 21.94% of interviewers or to 13.77% of respondents. Higher error risk was observed in complicated questions, such as questions assessing food quantification or referring to seasonally supplied food groups. After correcting the errors, the means and standard deviations of estimated food intakes all decreased. These findings suggested that interviewer error should not be ignored within face-to-face food frequency questionnaires and that more efforts are needed to monitor error-prone interviewers and respondents and reduce survey burdens in questionnaire design.

audio recording; data quality; food frequency questionnaire; interviewer error; multivariate outlier detection

Abbreviations: CMEC, China Multi-Ethnic Cohort; FFQ, food frequency questionnaire.

In large-scale epidemiologic studies, dietary measurements are commonly conducted through face-to-face interviews based on a food frequency questionnaire (FFQ) (1), particularly in low- and middle-income countries. Despite the increasing use of telephone- or web-based survey methods, face-to-face interviews remain a popular data collection method because of several advantages over other methods (2). In face-to-face interviews, interviewers can provide timely explanations for questions that confuse the respondents and confirm answers with the respondents if necessary, helping improve the data quality (3). In addition, the respondents are typically not required to use electronic products during the interviews, making it appropriate for the elderly or less-educated people (4). The ease of response makes

face-to-face surveys frequently used among populations in low- and middle-income countries (5, 6).

Interviewers can help improve the quality of survey data but may also introduce errors. In face-to-face interviews, the interviewers are expected to read the questions correctly, probe inadequate answers, and record answers accurately (7). During the multitasking procedure, the interviewers may intentionally or unintentionally fail to follow the survey guidelines, which is defined as interviewer error (8). Several studies have identified the presence of interviewer error. For instance, Hicks et al. (9) reported the errors introduced by the interviewers who significantly altered the question meanings or failed to ask for more information when the answers did not fit the option categories. The situation becomes more

severe if interviewers consciously manipulate the answers to shorten the interview time (10).

Interviewer error could be dire within face-to-face FFQs conducted in large multisite epidemiologic studies. First, dietary measurement via FFQ could be more error-prone than other epidemiologic exposures. As summarized by Willett and Lenart (11), FFQs are prone to errors, partly due to matching foods to a fixed group list and partly due to the cognitive challenge of assessing the frequency and amount of food consumed over a long time. Second, training and managing interviewers could become more difficult as the sample size and the number of survey sites increase (12), which may also increase the risk of interviewer error. Interviewer error is known to affect the data quality and even bias the results of epidemiologic studies (13). However, little is known about interviewer error within face-to-face FFQs, particularly in large multisite epidemiologic studies.

This study focused on evaluating interviewer error in the dietary data collection process in a large multisite epidemiologic study. Specifically, we aimed to 1) examine the source and magnitude of interviewer error; 2) capture the characteristics of interviewer error at the levels of interviewers, respondents, and questions; and 3) evaluate the impact of interviewer error on survey estimates. These findings may help identify potential problems within interviewer management and questionnaire design, and ultimately contribute to reducing interviewer error. These goals were achieved using dietary data from the China Multi-Ethnic Cohort (CMEC) study (14).

METHODS

We evaluated interviewer error based on dietary data from the CMEC study (14). The CMEC collected dietary data with a tablet-based FFQ through face-to-face interviews, and all interviews were audio recorded, allowing us to review the interviews and evaluate the interviewers' behaviors. In practice, because adequate quality control measures were taken in the CMEC and the selection of a random sample might be inefficient for understanding interviewer error, we searched for an error-prone sample for review. An overview of the interviewer error evaluation procedure is illustrated in Figure 1.

Baseline survey

The baseline survey of the CMEC was conducted in Southwest China between May 2018 and September 2019 (14). A total of 21,662 participants aged 30–79 years, recruited from 36 communities of Chengdu City, were included in this study. Dietary habits over the previous year were measured using an electronic FFQ covering the 13 major food groups in China: rice, wheat products, coarse grain, tubers, meat, poultry, seafood, eggs, fresh vegetables, soybean products, preserved vegetables, fresh fruits, and dairy products. The consumption of each food group was enquired about in 3 steps. The interviewers first asked the participants to report whether they consumed a certain food group (yes/no). If they responded yes, then they were required to recall how many times per day/week/month/year (frequency-related question) and how much each food was consumed at each meal (amount-related question). During the interviews, some standardized models (e.g., bowls and cups) were used to assist food quantification.

The baseline survey collected information using an interviewer-administered electronic questionnaire. During the data collection process, various quality control measures were taken to ensure data quality, such as recruiting interviewers with professional backgrounds, establishing standardized interviewing guidelines, and providing intensive interviewer training accordingly. In interviews, interviewers launched the electronic questionnaire by opening a cohort-developed application on the tablet (details are given in Web Appendix 1 and Web Figures 1–4, available at https://doi.org/10.1093/aje/kwac024) and were required to 1) read the questions correctly, 2) provide adequate neutral probing if necessary, and 3) interpret and enter the participants' responses correctly. In addition, the interviews were automatically audio recorded by the application. At the end of the working day, the survey data and the audio recordings were uploaded to a cohort-developed online system (see Web Appendix 1), allowing for retrospective access.

Error-prone sample identification

Among the 21,662 participants, 544 participants with missing data were excluded and 21,118 participants were included in the procedure of error-prone sample identification. We identified the error-prone sample in 2 steps. We first calculated the daily intakes of the 13 food groups by multiplying the frequency by the amount of food consumed and then applied a multivariate outlier-detection method (15) to detect the potential error-prone interviews. Details of the outlier detection method are provided below.

Outlier detection methods can help identify data points that deviate abnormally from the majority of the data. The most frequently used approach for multivariate outlier detection is the Mahalanobis distance (16), in which an observation is declared as an outlier if its distance exceeds a predefined threshold. Let $\mathbf{X}$ be a $n \times p$ matrix of p variables for n observations. The Mahalanobis distance for every observation x_i is then defined as $MD_i = \sqrt{(x_i - \hat{\mu})^T \hat{\Sigma}^{-1} (x_i - \hat{\mu})}$, with $\hat{\mu}$ and $\hat{\Sigma}$ representing the estimated mean and covariance matrix, respectively. We obtained robust estimates using the minimum covariance determinant (17–19). To determine the threshold, Filzmoser and others pointed out that an adjusted quantile was a better choice than the classical user-defined quantile, with the latter tending to provide a higher false classification rate (20, 21). Following Filzmoser, the adjusted quantile was applied in this study.

Audio review

Quality evaluation specialists reviewed the audio recordings attached to the error-prone sample. For every error-prone
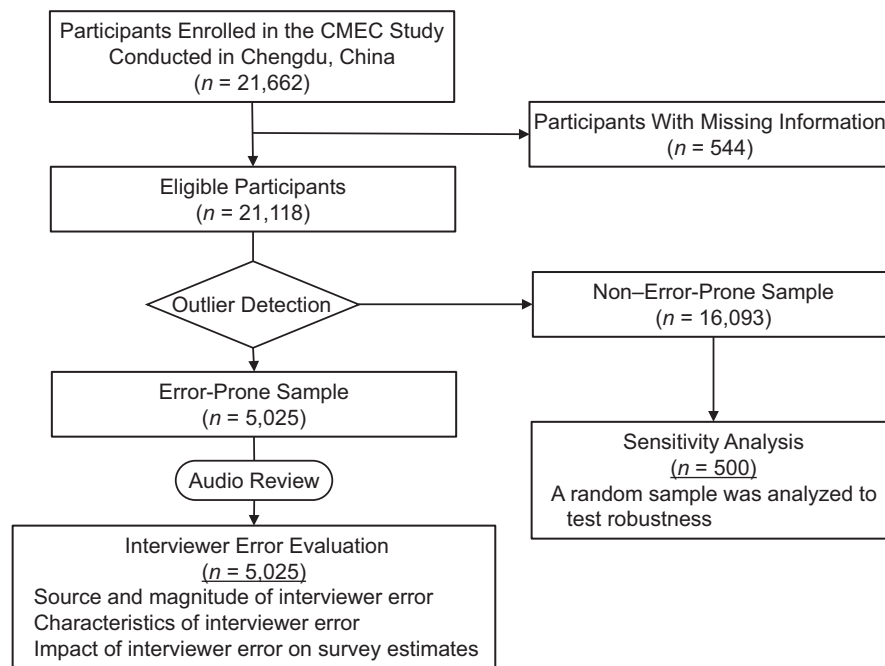
Figure 1. An overview of the interviewer error evaluation procedure and the findings from a large multisite study, the China Multi-Ethnic Cohort (CMEC), 2018–2019. FFQ, food frequency questionnaire.

interview, the specialist listened to the audio recording, compared what they heard with what the interviewer had keyed, and finally marked each question to indicate the presence and type of interviewer error. Every question was marked as one of the following types: no error, reading deviation, falsification, coding error, or unverifiable. In addition, any question identified with interviewer error, namely, falsification, reading deviation, or coding error, was tagged as a faulty question. The detailed marking scheme is displayed in Table 1.

Statistical analysis

Source and magnitude of interviewer error.   To examine the source and magnitude of interviewer error, we calculated the overall and specific error rates in terms of different error types. The error rate was defined as the number of faulty questions divided by the number of verifiable questions.

Characteristics of interviewer error.   The characteristics of interviewer error were analyzed at 3 levels, including the interviewer, respondent, and question levels. Notably, the 3 levels were hierarchically structured: One interviewer typically conducted interviews with a large number of respondents, and one respondent typically answered a certain number of questions. At the interviewer level, we analyzed the distribution and clustering tendency of interviewer error. To describe the distribution, we calculated the error rate of each interviewer and then displayed its distribution among the interviewers. To test whether the interviewer

error clustered towards some interviewers, we calculated the individual proportion each interviewer contributed to the total errors. We performed a standardized procedure to estimate the contribution proportion, given that the number of questions each interviewer asked could be different and a larger question number generally comes with more faulty questions and greater error contribution. Specifically, if we let er_i be the error rate of the ith interviewer and assume the question number asked by the interviewers is equal to n, the error contribution of the ith interviewer, then, could be expressed as a proportion, either $n \cdot er_i / \sum_j^m n \cdot er_j$ or $er_i / \sum_j^m er_j$. Similar to the analysis at the interviewer level, we calculated the error rate at the respondent level, and then displayed the distribution and clustering tendency of interviewer error among all respondents. At the question level, we examined the source and magnitude of interviewer error in terms of the 3 question types and 13 food groups.

Impact of interviewer error.   To quantify the impact of interviewer error on dietary assessment, we recoded all faulty questions and contrasted the daily intakes estimated between the raw and recoded data sets. Specifically, questions identified with coding error were replaced by the correct values, while others identified with reading deviation and falsification were set to missing values because the correct values were unknown.

Sensitivity analysis.   A random sample from the interviews not identified in the error-prone sample was selected and analyzed, to test whether the result from the error-prone sample was robust. In addition, we examined the changes

Table 1. The Marking Scheme Used in the Audio Review Process, the China Multi-Ethnic Cohort, 2018–2019

Status	Definition
No error	The interviewer delivers the question and options correctly, and keys the respondent's answer correctly.
Reading deviation	The interviewer delivers the question or options incorrectly or inappropriately. This error is coded when the interviewer misphrases the question or leads the respondent to elicit a certain answer.
Falsification	The interviewer falsifies the entire or part of the answer provided by the respondent. This error is coded when the interviewer fakes an answer without contacting the participant, or personally interprets an unclear response rather than asking for clarification.
Coding error	The interviewer keys the wrong code for the respondent's answer.
Unverifiable	The interview is unverifiable due to the background noise, low speaking volume, or the absence of audio files.

in daily intake by only correcting the questions with coding error.

RESULTS

Source and magnitude of interviewer error

The minimum covariance determinant approach identified an error-prone sample containing 5,025 error-prone interviews. The audio review results derived from the error-prone sample are described in Table 2. A total of 180,811 questions for 5,025 respondents and 351 interviewers were manually reviewed, with 6,799 (3.76%) questions being unverifiable and 174,012 (96.24%) questions verifiable. Among all verifiable questions, 13,855 (7.96%) were identified with interviewer error, with 5,200 having falsification, 4,393 with coding error, and 4,262 with reading deviation.

Characteristics of interviewer error

The distributions of interviewer error at the interviewer and respondent levels are characterized in Table 3 (details are given in Web Figures 5 and 6). At the interviewer level, all 351 interviewers were reviewed, and 345 (98.29%) were identified with at least 1 error. The error rates ranged from 0.00% to 56.19% among the interviewers, and half of them were lower than 6.36%. Of the 3 error types, the distribution

of falsification was significantly skewed to the left, showing a smaller median error rate (1.35%) and a larger mean error rate (3.34%) than reading deviation (median, 1.86%; mean = 2.43%) and coding error (median, 2.51%; mean = 2.76%). At the respondent level, similar to the findings at the interviewer level, we found that a large proportion of respondents were subject to interviewer error, and for most of the respondents the interviewer error rate was low. For instance, 3,704 (73.71%) respondents among all 5,025 respondents were identified with at least 1 interviewer error, and half of them had an error rate lower than 5.13%.

The clustering tendencies of interviewer error at the interviewer and respondent levels are displayed in Figure 2. At the interviewer level, we found that a large part of interviewer error tended to cluster among a small portion of interviewers, especially falsification. As shown in Figure 2A, among all 351 reviewers, the 4 (1.14%) most error-prone interviewers accounted for 6.78% of the total errors, and the 77 (21.94%) most error-prone interviewers contributed half of the total errors. Of the 3 error types, the clustering tendency was most obvious in falsification and least obvious in coding error. For instance, the 4 interviewers who were most prone to falsification accounted for 14.35% of falsification errors, while the 4 interviewers who were most prone to coding error accounted for 4.12% of coding errors.

Compared with the interviewer level, similar clustering tendency was observed at the respondent level. As shown in

Table 2. Overview of the Audio Review Results From an Error-Prone Sample, the China Multi-Ethnic Cohort, 2018–2019

Status	No.	Proportion, %
Unverifiable questions	6,799	3.76
Verifiable questions	174,012	96.24
No error	160,157	92.04
Reading deviation	4,262	2.45
Falsification	5,200	2.99
Coding error	4,393	2.52
Total	180,811	100.00

Table 3. Summary of Interviewer Error at the Interviewer and Respondent Levels, the China Multi-Ethnic Cohort, 2018–2019

Type of Error	Error Rate, %					
	Minimum	Lower Quantile	Median	Upper Quantile	Maximum	Mean (SD)
Interviewer level						
Overall	0.00	4.06	6.36	10.59	56.19	8.53 (7.82)
Reading deviation	0.00	0.92	1.86	3.33	25.93	2.43 (2.62)
Falsification	0.00	0.36	1.35	3.45	53.47	3.34 (6.13)
Coding error	0.00	1.64	2.51	3.52	11.11	2.76 (1.74)
Respondent level						
Overall	0.00	0.00	5.13	10.81	100.00	8.19 (10.94)
Reading deviation	0.00	0.00	0.00	2.86	40.91	2.52 (5.12)
Falsification	0.00	0.00	0.00	2.86	100.00	3.12 (8.24)
Coding error	0.00	0.00	2.56	3.33	100.00	2.55 (3.54)

Abbreviation: SD, standard deviation.

Figure 2B, a large portion of interviewer error was observed among a small number of respondents, especially falsification. Specifically, among all 5,025 respondents, the 50 (1.00%) respondents who were most prone to interviewer error contributed 8.38% of the total errors, and the 692 (13.77%) respondents who were most prone to interviewer error accounted for half of the total errors. Of the 3 error types, the 50 respondents who were most prone to falsification accounted for 19.82% of falsification errors, while the 50 respondents who were most prone to coding error accounted for 8.38% of coding errors.

We further characterized the interviewer error at the question level. The error rates of the 3 question types and 13 food groups as well as the proportions of different error types are depicted in Figure 3. The overall error rate was 7.96%, which mainly came from falsification (37.53%), coding error (31.71%), and reading deviation (30.76%).

Of the 3 question types, the most error-prone type was the amount-related question. The error rate of this question type reached 16.62%; namely, approximately 17 of 100 amount-related questions had interviewer error. The frequency-related question and the yes/no question were less error-prone, with error rates of 6.82% and 1.69%, respectively. Additionally, the structure of interviewer error varied by question type. Reading deviation generally dominated in the yes/no question, falsification dominated in the frequency-related question, and coding error dominated in the amount-related question.

Of the 13 food groups, the error rates ranged from 5.27% to 15.53%. The highest error rate was observed for soybean products (15.53%), followed by coarse grain (8.89%) and several seasonally available food groups, such as fresh vegetables (9.01%), fresh fruits (8.83%), and tubers (8.23%). The remaining food groups were less error-prone. Moreover, significant differences in terms of the structure of interviewer error were observed among the 13 food groups. Falsification was the most observed error type, observed in

8 of 13 food groups. Preserved vegetables had the highest proportion of falsification errors (56.41%). In addition, falsification was related to the order in which food groups were presented in the survey; namely, those listed at the end of the questionnaire were far more prone to falsification than groups listed at the beginning of the questionnaire. The highest proportions of coding error were observed for rice (47.51%), wheat products (48.98%), and soybean products (56.94%), while reading deviation dominated in coarse grain (38.93%) and meat (42.28%).

Among the 16,093 interviews not identified in the error-prone sample, a random sample of 500 interviews was selected and analyzed. For the characteristics of interviewer error, the random sample provided similar results as the error-prone sample (details are given in Web Appendix 2 and Web Table 1). For instance, the overall error rate of the random sample was 7.37%, slightly lower than that of the error-prone sample (7.96%). The similar results between the 2 samples indicated that the findings from the error-prone sample were robust.

Impact of interviewer error

A total of 13,855 faulty questions were recoded. Specifically, 4,393 (31.71%) responses with coding error were replaced by the correct values, while 9,462 (68.29%) responses with falsification and reading deviation were set to missing values. Figure 4 compares the differences in the means and standard deviations of daily food intake estimated from the raw data and the recoded data. Negative changes for both statistics appeared in all food groups after recoding. The relative reductions in mean daily food intake ranged from 1.41% (rice) to 35.88% (soybean products). More significant decreases were observed in the standard deviations of daily food intake, in which the relative reductions ranged from 3.07% (preserved vegetables) to 51.15% (eggs). Moreover, we found that nearly all the means and standard deviations
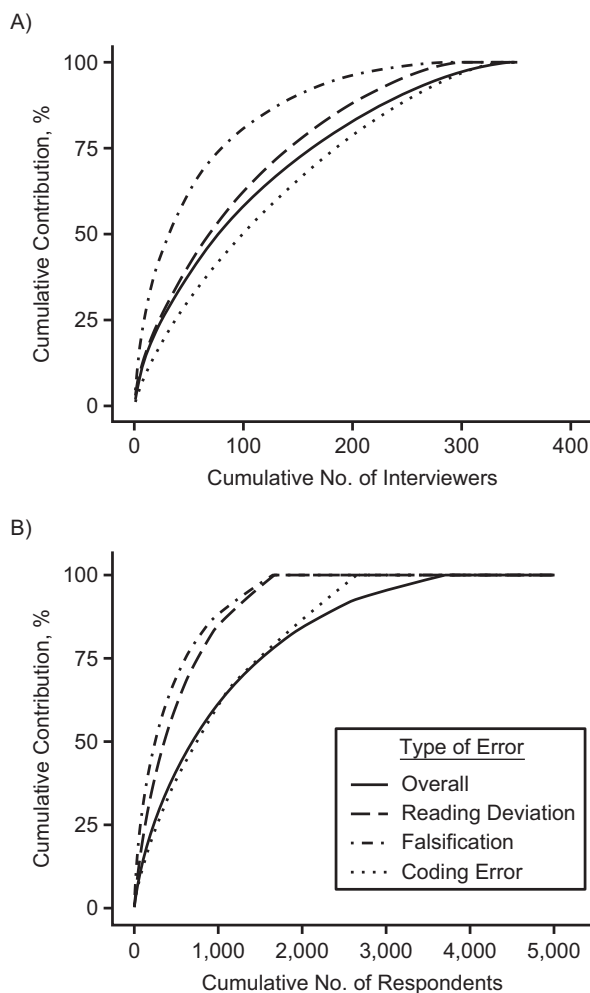
A)

B)

Figure 2. The clustering tendency of interviewer error analyzed at the interviewer level (A) and the respondent level (B), the China Multi-Ethnic Cohort, 2018–2019. Take the interviewer level as an example: first, we calculated the individual proportion each interviewer contributed to the total number of errors. It can be obtained by dividing the number of questions with error from this interviewer by the total number of questions with error from all interviewers. Then we calculated the cumulative contribution proportions of the interviewers, in which the interviewers were ordered from highest to lowest contributions. The closer the line to the top-left corner, the more significant the clustering tendency. The clustering tendency at the respondent level is generated in the same way as that at the interviewer level.

of the food intakes of 13 food groups decreased after only correcting the coding error, except for coarse grain.

DISCUSSION

Interviewer error may directly affect data quality and even lead to biased results (13, 22). This study aimed to evaluate the interviewer error within a face-to-face FFQ in a large multisite epidemiologic study. Based on the baseline survey of the CMEC, we reviewed the error-prone part of the field

interviews and found that only approximately one-quarter of them did not have interviewer error. The interviewer error was derived primarily from falsification, followed by coding error and reading deviation. Moreover, the prevalence of interviewer error varied greatly by interviewer, respondent, and question.

Our findings suggested that the interviewers and respondents were subject to interviewer error. Nearly all interviewers and three-quarters of respondents were affected by interviewer error. Fortunately, for most interviewers and respondents, the interviewer error rates were low, and a large part of interviewer error could be attributed to a small number of interviewers and respondents who were more prone to interviewer error than others. For example, half of the interviewer error could be attributed to 21.94% of interviewers and 13.77% of respondents. According to these findings, focusing on some of the most error-prone interviewers or respondents could help avoid the majority of interviewer error, especially falsification.

In addition, the question characteristics play a role in making interviewer error more or less likely. Generally, complicated questions may place an additional burden on both interviewers and respondents, increasing the risk of interviewer error (23). Of the 3 question types, the questions dealing with food quantification appeared the most problematic, consistent with previous studies (24, 25). Of the 13 food groups, the food groups whose quantification is complex and those representing a large number of categories or appearing seasonally tended to be more error-prone. For soybean products, the most error-prone food group, its quantification additionally required the conversion of the weight of processed soybean products (e.g., tofu and soybean milk) into the weight of raw soybeans. The error rates of several seasonal food groups, such as tubers, fresh vegetables, and fruits, all exceeded the average level. The interviewers likely failed to capture the seasonal variation in food supply and consumption (26, 27). The identification of problem questions provides opportunities for future advances in questionnaire design. For instance, including both the pre- and post-harvest seasons when designing questions will help reduce interviewer error, but so will including appropriate questions to reduce the burden of doing "mental math," particularly in the case of soybean products.

Researchers could be at risk of overestimating dietary intake if interviewer error is ignored. After error correction, both the means and standard deviations of daily intakes for all food groups decreased. Moreover, although a total of 13,855 errors were identified in this study, only one-third of them were able to be replaced with correct values, while others were set to missing. This process could cause a large loss of information, which was unavoidable given the interviewer-error evaluation approach applied in this study. However, our primary goal was to examine the prevalence of interviewer error, and the correction was an additional outcome.

Measures for regulating interviewer behaviors in interviewer-administered surveys have long been implemented to improve data quality (7, 28). In the CMEC, common quality control measures have been conducted to ensure data quality, but the errors introduced by interviewers still
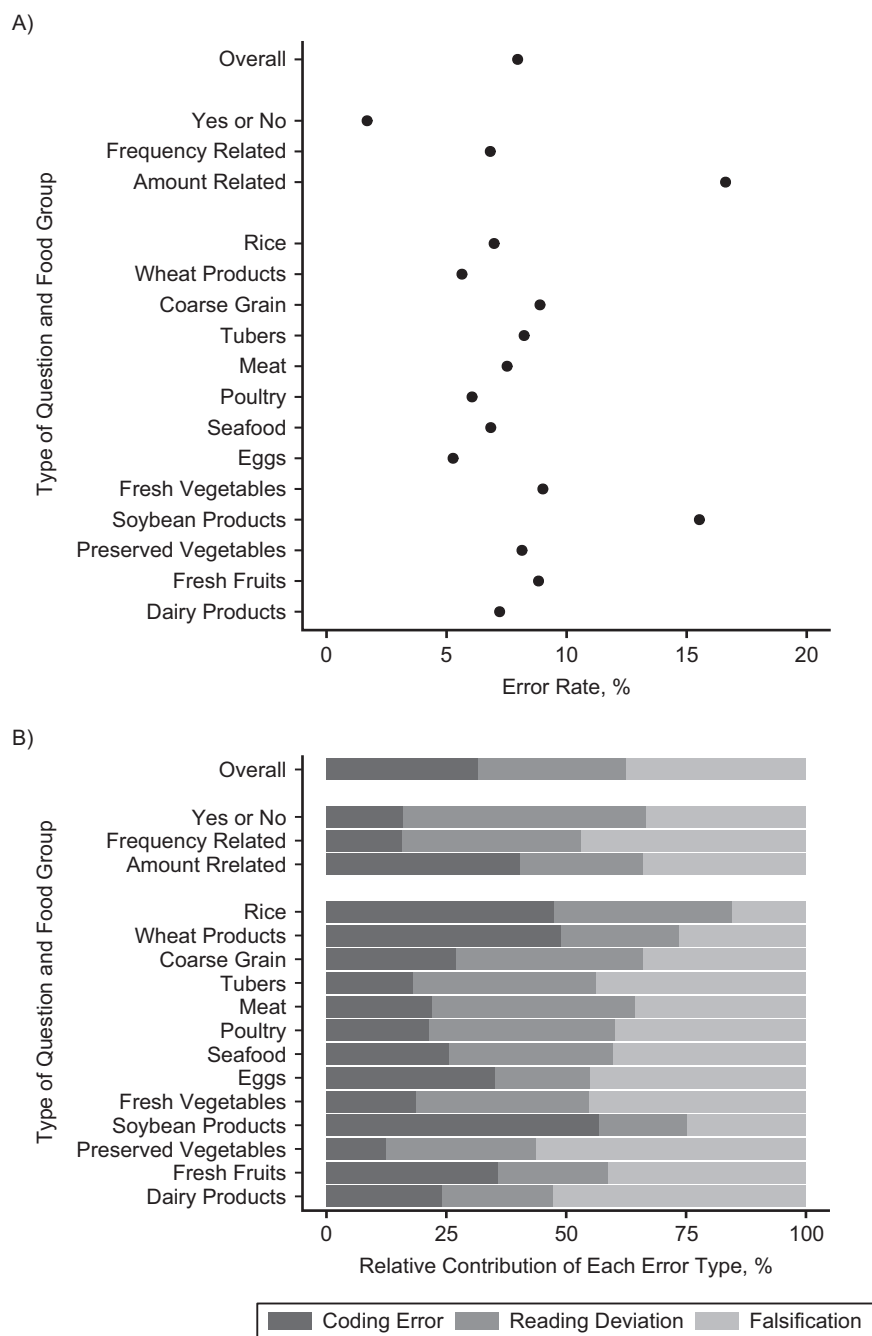
Figure 3. The overall, question-specific, and food-specific error rates and relative contributions of 3 error types, the China Multi-Ethnic Cohort, 2018–2019. Food groups were roughly ordered according to the order in the questionnaire.

existed in approximately three-quarters of all error-prone interviews. This finding indicated that the routine quality control measures used in epidemiologic studies are valuable but insufficient and should not be substituted for the monitoring of data for quality.

Some limitations need to be considered. First, although interviewer error can occur in several ways (29), this study focused on only error arising when the interviewers fail to

follow the survey guidelines, namely, fail to read the questions correctly, inadequately probe the details, and record answers inaccurately. We did not consider error associated with reactivity, which arises when respondents provide answers they believe the interviewer wants to hear in the face-to-face survey, as the data regarding reactivity is hard to collect. Second, our evaluation of interviewer error relied on audio recordings and, thus, may miss the unspoken part of
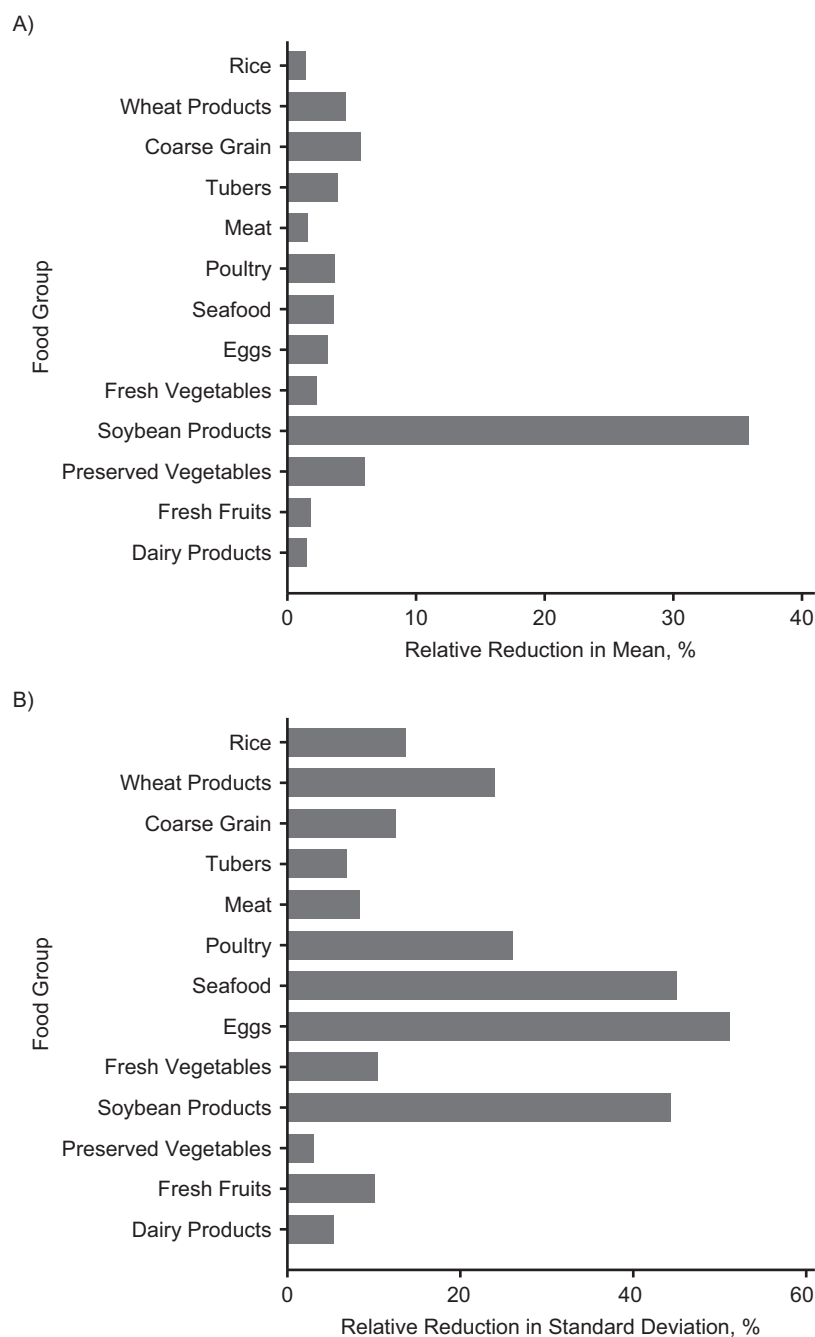
A)

B)

Figure 4. The relative reductions in the means (A) and the standard deviations (B) of daily food intake resulting from the interviewer error correction, the China Multi-Ethnic Cohort, 2018–2019.

communication. Finally, the findings of the present study were obtained from face-to-face FFQs, so the findings, especially those associated with questionnaire design, should be generalized to other interviewer-administered questionnaires only with care.

In conclusion, interviewer error is a nonignorable problem of the face-to-face FFQ in large multisite epidemiologic studies. Our results suggested that the leading source of interviewer error was falsification, followed by coding error and reading deviation. Interviewer error was common among the interviewers and respondents, and a large portion of error was attributed to a small number of interviewers or respondents. It is necessary to strengthen the management and monitoring of the most error-prone interviewers or respondents. Additionally, complicated questions, such as those assessing food quantification or referring to seasonally supplied food groups, may place additional burdens on the interviewers and respondents, thus increasing the risk of

interviewer error. Therefore, more efforts are needed to reduce the survey burdens in questionnaire design.

REFERENCES

1. Shim JS, Oh K, Kim HC. Dietary assessment methods in epidemiologic studies. *Epidemiol Health.* 2014;36:e2014009.
2. Couper MP. The future of modes of data collection. *Public Opin Q.* 2011;75(5):889–908.
3. Irvine A, Drew P, Sainsbury R. "Am I not answering your questions properly?" clarification, adequacy and responsiveness in semi-structured telephone and face-to-face interviews. *Qual Res.* 2013;13(1):87–106.
4. de Leeuw ED. Choosing the method of data collection. In: de Leeuw ED, Hox JJ, Dillman DA, eds. *International Handbook of Survey Methodology*. New York, NY: Larence Erlbaum Associates; 2008:113–135.
5. Chen Z, Lee L, Chen J, et al. Cohort profile: the Kadoorie Study of Chronic Disease in China (KSCDC). *Int J Epidemiol.* 2005;34(6):1243–1249.
6. Pisa PT, Landais E, Margetts B, et al. Inventory on the dietary assessment tools available and needed in Africa: a prerequisite for setting up a common methodological research infrastructure for nutritional surveillance, research, and prevention of diet-related non-communicable diseases. *Crit Rev Food Sci Nutr.* 2018;58(1):37–61.
7. Fowler F, Mangione T. *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. Vol. 18. Newbury Park, CA: SAGE; 1990.
8. Davis RE, Couper MP, Janz NK, et al. Interviewer effects in public health surveys. *Health Educ Res.* 2010;25(1):14–26.
9. Hicks WD, Edwards B, Tourangeau K, et al. Using CARI tools to understand measurement error. *Public Opin Q.* 2010; 74(5):985–1003.
10. Murphy J, Biemer P, Stringer C, et al. Interviewer falsification: current and best practices for prevention, detection, and mitigation. *Stat J IAOS.* 2016;32(3):313–326.
11. Willett W, Lenart E. Reproducibility and validity of food frequency questionnaires. In: Willett W, ed. *Nutritional Epidemiology*. 3rd ed. New York, NY: Oxford University Press; 2013:96–141.
12. Ackermann-Piek D, Silber H, Daikeler J, et al. Interviewer training guidelines of multinational survey programs: a total survey error. *Perspectives.* 2020;14(1).
13. West BT, Blom AG. Explaining interviewer effects: a research synthesis. *J Surv Stat Methodol.* 2017;5(2):175–211.
14. Zhao X, Hong F, Yin J, et al. Cohort profile: the China Multi-Ethnic Cohort (CMEC) study. *Int J Epidemiol.* 2021; 50(3):721–721l.
15. Rousseeuw PJ, Hubert M. Anomaly detection by robust statistics. *Wiley Interdiscip Rev Data Min Knowl Discov.* 2018;8(2):e1236.
16. Mahalanobis PC. On the generalized distance in statistics. *Proc Nat Inst Sci India.* 1936;2(1):49–55.
17. Rousseeuw PJ. Multivariate estimation with high breakdown point. In: Grossman W, Pflug I, Wertz W, eds. *Mathematical Statistics and Applications*. Dordrecht, the Netherlands: Reidel Publishing Co.; 1985:283–297.
18. Rousseeuw PJ. Least median of squares regression. *J Am Stat Assoc.* 1984;79(388):871–880.
19. Rousseeuw PJ, Driessen KV. A fast algorithm for the minimum covariance determinant estimator. *Dent Tech.* 1999; 41(3):212–223.
20. Filzmoser P, Garrett RG, Reimann C. Multivariate outlier detection in exploration geochemistry. *Comput Geosci.* 2005; 31(5):579–587.
21. Cabana E, Lillo RE, Laniado H. Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. *Stat Pap (Berl).* 2019;62(4):1583–1609.
22. Beullens K, Loosveldt G. Interviewer effects in the European Social Survey. *Surv Res Methods.* 2016;10(2):103–118.
23. Japec L. Interviewer error and interviewer burden. In: Lepkowski JM, Tucker C, Brick JM, Leeuw EDD, Sangster RL, eds. *Advances in Telephone Survey Methodology*. Hoboken, NJ: John Wiley & Sons; 2007:187–211.
24. Rumpler WV, Kramer M, Rhodes DG, et al. Identifying sources of reporting error using measured food intake. *Eur J Clin Nutr.* 2008;62(4):544–552.
25. Souverein OW, de Boer WJ, Geelen A, et al. Uncertainty in intake due to portion size estimation in 24-hour recalls varies between food groups. *J Nutr.* 2011;141(7):1396–1401.
26. Stelmach-Mardas M, Kleiser C, Uzhova I, et al. Seasonality of food groups and total energy intake: a systematic review and meta-analysis. *Eur J Clin Nutr.* 2016;70(6):700–708.
27. Zhu Z, Wu C, Luo B, et al. The dietary intake and its features across four seasons in the Metropolis of China. *J Nutr Sci Vitaminol (Tokyo).* 2019;65(1):52–59.
28. Neta G, Samet JM, Rajaraman P. Quality control and good epidemiological practice. In: Ahrens W, Pigeot I, eds. *Handbook of Epidemiology*. New York, NY: Springer; 2014: 525–576.
29. Groves RM. *The Interviewer as a Source of Survey Measurement Error. Survey Errors and Survey Costs*. New York, NY: John Wiley & Sons; 1989:357–406.