Keying Errors Caused by Unusual Keypunch Codes: Evidence from A Current Population Survey Test

Harley Frazis, Jay Stewart, Bureau of Labor Statistics
Jay Stewart, BLS, 2 Massachusetts Ave. NE, Rm. 4945Washington, DC 20212-0001

Key words: Computer assisted interviewing, interviewer errors.

I. Introduction

Over the last few years, computer assisted interviewing (CAI) has become the preferred way to conduct household interviews. CAI has two main advantages. First, it makes it possible to tailor the interview to the respondent by permitting fills and complicated skip patterns that would not be possible in a paper and pencil interview (PAPI). Second, data can be edited and (if necessary) corrected as it is entered. Despite these advantages, CAI introduces another source of error that was not present in PAPI: keying errors. Before CAI was introduced in the Current Population Survey (CPS), for example, responses were transformed into machine-readable form by scanning copies of the interview booklets. Moving from this method of transcription to CAI may introduce errors because interviewers may not be proficient in entering data into the computer.

Despite its importance as a potential source of non-sampling error, there has been little research on keying errors. In a recent study of computer-assisted personal interviewing (CAPI) interviewing, Dielman and Couper (1995) investigated the incidence of keying errors in the recording of responses to closed-ended questions. They compared the interviewer coded responses to tape recordings of the actual interviews. They found only 16 keying errors out of 16,778 questions, leading them to conclude that "[t]he miskeying of answers to closed-ended questions...does not appear to be a major cause for concern on carefully designed CAPI surveys."

This article looks at keying error that can occur when interviewers are presented with unusual keypunch codes — codes where the correct keying is different from what the interviewer is used to. Using data from a supplement to the Current Population Survey (CPS) we find that such categories can cause rates of miskeying orders of magnitude higher than found by Dielman and Couper (1995). We estimate that unusual response categories resulted in miskeying error rates of 12 percent. The lesson that we draw from this is that avoidance of such codes is part of careful survey design.

II. Data

The CPS is a monthly household survey of 60,000 households conducted by the US Census Bureau for the Bureau of Labor Statistics. In July 1995, a supplement to the main survey was administered to test questions on educational attainment. This supplement was administered to one-quarter of the sample. The mode of administration was mostly computer-assisted telephone interviewing (CATI), with some CAPI interviews.[1]

The CPS normally asks about educational attainment with the following question:

E1. What is the highest level of school ... has completed or the highest degree ... has received?

The supplement consisted of a series of follow-up questions to E1, where the specific question asked depended on the response to E1. The question we examine in this paper was asked of respondents who reported educational attainment of less than a high school diploma in E1. The purpose of this question was to see if these respondents had obtained the equivalent of a High School diploma by non-traditional means, but failed to report it in E1. The best-known way of obtaining such an equivalency in the United States is by passing the General Educational Development (GED) tests. These respondents were asked:

E2. Did you ever get a High School diploma by completing High School OR through a GED or other equivalent?

---

[1]The CPS is administered by address. Addresses are in the sample for four months, out for eight months, and back in sample for four months. The supplement was administered to households in their fourth and eighth month in the survey, who are normally contacted by telephone.

1) Yes, completed High School.
2) Yes, GED or other equivalent.
3) No.

As can be seen, the response categories include two "Yes" categories instead of one. The usual CPS format for recording answers to yes-no questions is for interviewers to key '1' for yes and '2' for no.

This question generated a very high estimate of the number of high school completers who did report having a high school diploma in E1. Based on the responses to E1, we estimated that 44.5 million persons aged 15+ had less than a high school degree. However, based on the responses to E2, 5.7 million (12.8 percent) of these had GEDs (response 2), and an additional 2.0 million (4.5 percent) had completed high school (response 1).[2] We believe that the '2' responses were mostly spurious, and that they were caused by miskeying (we believe that the '1' responses were also largely spurious, but leave discussion of this to the footnotes).

The major piece of evidence that most of the GED responses were spurious is that an estimated 900 thousand 15- and 16-year-olds had GEDs, 12 percent of the total number of GEDs found in the supplement. Administrative data (GED Testing Service, 1992) show that less than 500,000 GEDs are awarded annually, and that only about 1 percent of these are awarded to 15-16 year-olds. Since it is highly unlikely that people in this age group received their GED when they were 14 or younger, we should expect to find no more than 5-10,000 (allowing for sampling error) GEDs in this age group. Hence, virtually all of these '2' responses must be erroneous.[3]

III. Explaining the Error

The explanation of this high error rate appears to be keying errors. The response codes (1=Yes, 2=Yes, 3=No) seem to have confused some interviewers used to the usual (1=Yes, 2=No) coding. While monitoring interviews, one author (Frazis) observed a "No" response coded as a '2'. This pointed us toward keying errors as a possible explanation for the high error rate.

It is possible that most of the erroneous responses were due to respondents' misunderstanding the question. To distinguish between respondent misunderstanding and miskeying, we look at interviewer effects. We reason that the identity of the interviewer should have a much larger effect on responses if the problem is miskeying than if the problem is respondent understanding (interviewers are working from a standard CATI script). To test for interviewer effects, we fit a model that estimates the dispersion in interviewer probabilities of getting response '2'. That is, we assume that each interviewer has his or her own probability of keying in a '2' response, and that these probabilities are distributed beta across interviewers. We use the maximum-likelihood estimates of the parameters of the beta distribution to calculate the standard deviation of the probability of getting a '2' response across interviewers. (See Kleinman 1973 for a more detailed explanation of this 'beta-binomial' model.)

There may be some dispersion in interviewer probabilities caused by non-random assignment of cases among interviewers. Within CATI sites, households are assigned at random, but cases are distributed among the three CATI sites geographically, as are all CAPI cases. There may also be some correlation of responses within households. We therefore need to compute a measure of the dispersion of interviewer probabilities for a baseline case. To do this, we estimated the dispersion in interviewer probabilities for a similar question, E3. This question was asked of those who responded high school graduate to E1:

E3. How did you get your High School diploma?
1) Graduation from High School
2) GED or other equivalent.

The proportion who answered "GED" in E3 was similar to E2 (11.0 percent on a weighted basis, 10.9 percent unweighted). The sample size – the number of interviewers with any valid responses to E3 – was 993. We estimated the standard deviation of interviewer probabilities of getting response '2' to E3 to be 5.4 percent (standard error 0.4 percent).

---

[2] On an unweighted basis, out of 5519 respondents who initially reported less than a high school degree, 672 (12.2 percent) were recorded as GEDs and 244 (4.4 percent) were recorded as having completed high school.

[3] We also found that a large number of 15-16 year-olds responded '1' to E2, indicating that they had completed high school. We believe that these respondents may have understood the question to be asking whether they *would* complete high school, not whether they *had* completed high school.

In contrast, the estimated standard deviation of interviewer probabilities of getting response '2' to E2 was much larger, 12.6 percent with a standard error of 0.7 percent.[4] (The sample size for E2 was 951.) This confirms what we found when "eyeballing" the data: some interviewers had high proportions of '2' responses (one had 13 '2' responses out of 15).[5] (Figures 1 and 2 show the distribution of proportions responding '2' across interviewers with 10 or more responses, for E2 and E3.) We conclude from this that it is differing probabilities of making keying errors, rather than non-random assignment of cases, that explain the high dispersion of '2' responses to E2 across interviewers.[6]

IV. Conclusion

To help gauge the potential effects of these keying errors, we computed the number of people with GEDs using raw data from question E2 and using corrected data from E2. Assuming that all of the '2' responses to E2 for 15-16 year-olds are erroneous, the error rate is approximately 12.1 percent. This means that for persons aged 17+ the percent of additional people with GEDs is closer to 1 percent of people who initially reported no high school diploma than to the 13 percent actually picked up by E2. This translates into 400,000 additional GEDs rather than 4.8 million. (We remind the reader that the supplement was a test and that these numbers did not affect official statistics.)

This example shows the dangers of using keypunch codes that are similar, but not identical, to familiar patterns. In this case, the incidence of keying errors appears to have been large enough to inflate the number of GEDs in the population asked E2 by an order of magnitude. To correct this problem, we proposed changing the coding from 1,2,3 to 4,5,6. However, the question was dropped from the sequence, so we are unable to report whether this solved the problem. Whether or not this would have solved the problem, it is clear that survey designers should avoid as much as possible using questions with such unusual keypunch codes.

Disclaimer

The opinions expressed here are those of the authors and do not necessarily reflect the views of the Bureau of Labor Statistics or those of other staff members. We would like to thank Mick Couper, Clyde Tucker, and colleagues at BLS and the Census Bureau for helpful comments.

References

Dielman, Lynn, and Couper, Mick P. (1995). "Data quality in a CAPI Survey: Keying Errors." *Journal of Official Statistics* 11:2, 141-46.

GED Testing Service. (1992) The 1991 Statistical Report. Washington, DC: American Council on Education.

Kleinman, J. C. (1973). "Proportions With Extraneous Variance: Single and Independent Samples." *Journal of the American Statistical Association* 68, 46-54.

---

[4] For purposes of estimation, we do not distinguish between responses '1' and '3' to E2.

[5] Removal of this observation had essentially no effect on the results..

[6] As a check, we estimated the 'beta-binomial' for the '1' responses in E2. The results were similar to those for E3, indicating that the spurious '1' responses were not due to miskeying. We suspect that the spurious '1' responses were due to respondents misunderstanding the question. Note that it is possible that some of the spurious responses to E2 resulted from misunderstanding the question, but the interviewer effects point to keying errors as the primary cause.

Figure 1: Density of Interviewers' Percentages of '2' Responses to Question E2 (Interviewers with at least 10 responses to E2)

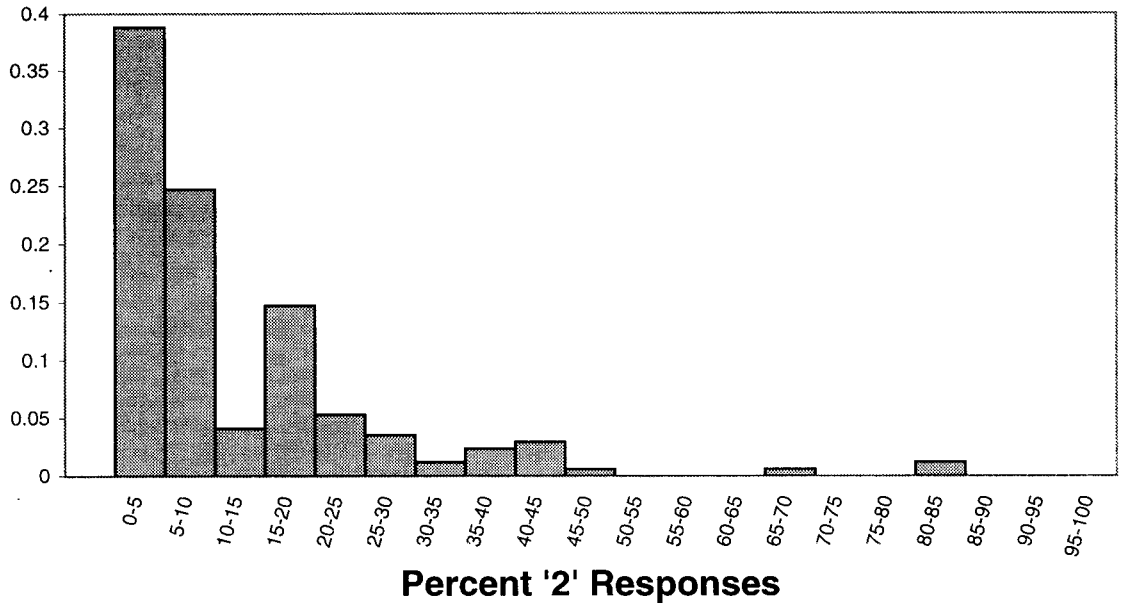

Figure 2: Density of Interviewers' Percentages of '2' Responses to Question E3 (Interviewers with at least 10 responses to E3)