Reconciling educational attainment questions in the CPS and the census

David A. Jaeger

The decennial census and the Current Population Survey (CPS) are two of the most widely used household surveys of the U.S. population. Both surveys measure the educational attainment of their respondents. From the 1940s to the 1990s, both surveys asked the respondents to report their *highest grade attended* and whether they had completed that grade, permitting the creation of the measure *highest grade completed.* Beginning with the 1990 census and the January 1992 CPS, both surveys have asked instead about individuals' *highest degree received.* An earlier research summary in the *Review* documented both the motivation for the change and the specific wording of the old and new questions.[1]

Researchers who previously used the continuous variable *highest grade completed* created from responses to the questions on education in the CPS and the census must now find new ways to represent educational attainment. Two possible methods of doing so are to *impute* a measure of the highest grade completed, using responses to the new question, and to *recode* responses to both questions, to provide comparable aggregate categories.[2] This report presents a summary account of these methods. The proposed methods are most useful to researchers using CPS and census microdata who wish to bridge the break in the question, but they also provide a benchmark for the magnitude of the differences in aggregate measurements of educational attainment using the old and the new questions. In general, the measure of the highest grade completed imputed from the new question is comparable to the highest grade completed as measured by the old question. And the overall average imputed highest grade completed is essentially the same as the average actual highest grade completed. However, the imputed highest grade completed overstates the highest grade completed for individuals at the lowest end of the educational distribution and slightly understates the highest grade completed for individuals at the highest end of the distribution. In addition, one can reconcile the old and new questions into four educational categories and make the distributions of educational attainment almost identical with the old and new questions.

Data

The data used were a matched sample of the 1991 and 1992 March CPS's. The sample design of the CPS allows one to match a portion of individuals in the sample across those years and create a new sample that contains responses to both the old and new CPS questions.[3] Individuals in rotation groups one through four of the March 1991 CPS were asked the old question on educational attainment in 1991. These individuals appear as rotation groups five through eight in the March 1992 CPS and were asked the new question in 1992. To reduce the likelihood that the sample contains individuals whose true level of schooling changed between 1991 and 1992, the sample is limited to individuals 25 to 64 years old who were not enrolled in school in either year.[4] The question asked in the 1990 census is essentially the same as the new question in the CPS. The census question provides more detail for very low levels of education (that is, "no school completed," "kindergarten," and "nursery school" in the census, compared with "less than 1st grade" in the CPS) and combines "5th or 6th grade" and "7th or 8th grade" into one category. Table 1 lists the response categories for the new question in both surveys, along with the codes used in the microdata.

Imputing "highest grade completed"

Given a data set with responses to both questions, a natural choice for imputing values for the highest grade completed with regard to the new question is some measure of central tendency, such as the mean or median of the highest grade completed from the old question, conditional on each value of the new question. Now, while conditional means are an obvious choice, they are sensitive to outliers. Medians are less sensitive to outliers, and modes provide the greatest number of observations whose imputed value is the same as the observed highest grade completed. In addition to describing the categories and listing the codes used in the microdata files, table 1 shows the mean, median, and modal values of the highest grade completed, conditional on each value of the new question. Note that the median and modal values are the same for each level.

To illustrate how well these imputed values compare with the observed highest grade completed, chart 1 plots the average imputed highest grade completed for each of the 19 values of the actual highest grade completed. The solid 45-degree line indicates perfect agreement between the actual and average imputed measures. The chart also shows the distribution of the sample across the values. Clearly, the imputed values overstate the average highest grade completed for individuals at the lower end of the educational distribution and understate somewhat the average highest grade completed for individuals who, in fact, completed 2 or 3 years of education after the 12th grade.

To address this issue, table 1 also shows "assigned" imputed values of the

David A. Jaeger is a former research economist in the Office of Employment Research and Program Development, Bureau of Labor Statistics. He is currently an associate professor, Department of Economics, Hunter College and the City University of New York Graduate School.

Table 1. Imputations of highest grade completed for new education questions

New question category	Codes		Imputation method		
	CPS	Census	Mean	Median and mode	Assigned
No school completed	...	01	...	...	0
Nursery school	...	02	...	...	0
Kindergarten	...	03	...	...	0
Less than 1st grade	31	...	1.30	0	0
First, 2nd, 3rd, or 4th grade	32	04	3.92	3	2.5
Fifth or 6th grade	33	...	6.22	6	5.5
Seventh or 8th grade	34	...	7.84	8	7.5
Fifth, 6th, 7th, or 8th grade	...	05	7.36	8	6.5
Ninth grade	35	06	9.08	9	9
Tenth grade	36	07	9.90	10	10
Eleventh grade	37	08	10.81	11	11
Twelfth grade, no diploma	38	09	11.38	12	12
High school graduate (high school diploma or equivalent)	39	10	12.00	12	12
Some college, but no degree	40	11	13.35	13	13
Occupational/vocational associate's degree in college	41	12	13.87	14	14
Academic associate's degree in college	42	13	14.29	14	14
Bachelor's degree (for example, B.A., A.B., B.S.)	43	14	16.04	16	16
Master's degree (for example, M.A., M.S., M.Eng., M.Ed., M.S.W., M.B.A.)	44	15	17.57	18	18
Professional school degree (for example, M.D., D.D.S., D.V.M., L.L.B., J.D.)	45	16	17.71	18	18
Doctoral degree (for example, Ph.D., Ed.D.)	46	17	17.84	18	18

NOTE: Data tabulated from a matched sample of individuals 25 to 64 years old from the March 1991 and 1992 Current Population Surveys. Reproduced in part from David A. Jaeger, "Reconciling the Old and New Census Bureau Education Questions: Recommendations for Researchers," *Journal of Business and Economic Statistics*, July 1997, pp. 300–09.

highest grade completed. The assignments are the same as the median and modal values for each category of grades, except for first through fourth grades, fifth and sixth grades, and seventh and eighth grades, each of which category is assigned a value at the midpoint of its range of values. Average imputed values for the "assigned" imputation method are also plotted in chart 1, which shows that this method comes closer, on average, to the actual highest grade completed at the low end of the educational distribution than do either of the other imputation methods. The chart also shows that very few in the sample stopped their education at less than the eighth grade, however, so the correction may not matter much in practical terms. The average imputed highest grade completed for the entire sample is essentially the same as the average of 13.00 for the actual highest grade completed, using any of the methods.

Creating comparable categories

In addition to using a measure of the highest grade completed, many researchers are interested in grouping individuals by more aggregated educational attainment categories. The four categories most often used are high school dropouts, high school graduates, individuals with some college, and college graduates. In establishing a scheme to recode the old and new questions into these categories, the aim is to preserve the rough similarity in the distribution of educational attainment between the two questions, while keeping the categories conceptually similar. However, because the old question provides no information about whether 12th graders did in fact graduate and receive a diploma, this article uses instead the classification "12th grade" rather than "high school graduate." The categorization that provides both a high degree of conceptual agreement between the questions and a high rate of matching between the recoded variables is presented in table 2.

For the most part, the coding scheme for the old question is as one might expect. The one exception is individuals who attended, but did not complete, 13th grade. These individuals are usually categorized as having a high school diploma, rather than having some college. The rate of matching is improved by including them among those with "some college," however, as 74.5 percent of individuals who attended, but did not complete, 13th grade report having "some college, but no degree" on the new question.

Researchers interested in decomposing the "college graduate" category into "college" and "postcollege" categories may group individuals who reported finishing 16th or 17th grade into a "4 or 5 years of college" category. Typically, researchers impute a 4-year college degree to individuals who report completing 16th grade. The match between the recoded old and new questions is improved, however, by also including in this category those who finished 17th grade. Among individuals who reported attending 17th grade with the old question, 57.9 percent held only a bachelor's degree under the new question, while 35.3 percent reported receiving a master's degree or higher. The majority of the remaining 6.8 percent report "some college, but no degree" or a "professional school degree."

The recoding scheme for the new question is also straightforward. Note that both of the categories "12th grade, no diploma" and "high school graduate" are recoded into "12th grade." This provides conceptual agreement between the new question and the old question, which could not separately identify those who received a high school diploma. In addition, the majority (55.3 percent) of individuals who reported completing 12th grade without receiving a diploma under the new question reported finishing 12th grade with the old question. Researchers interested in splitting the "college graduate" category may include master's, professional, and doctoral degree recipients in a "postcollege" category.

Table 3 shows the cross tabulation of the recoded old and new educational at-

Chart 1. Average imputed highest grade completed and share of sample, by actual highest grade

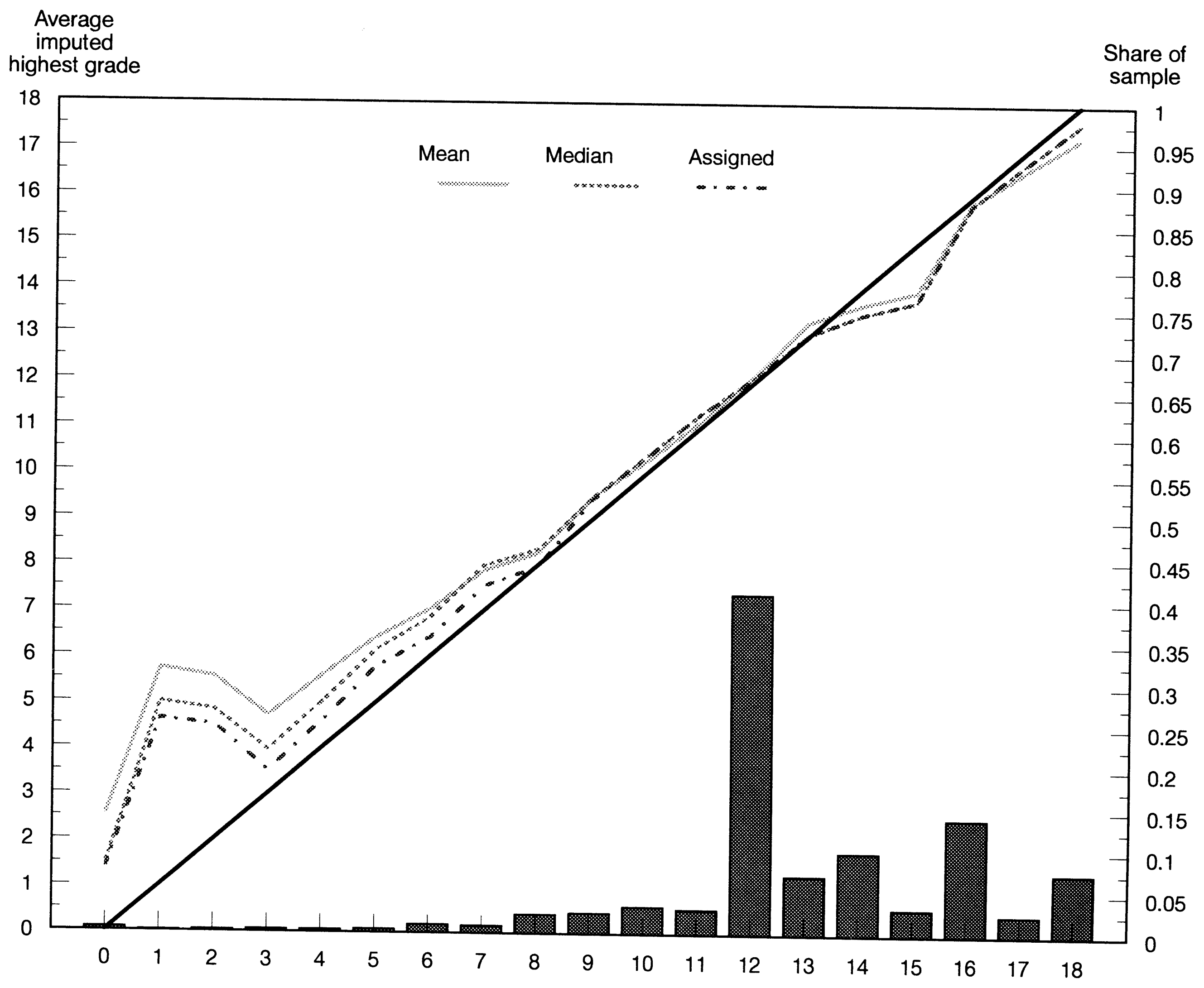


Table 2. Categorical recoding scheme for old and new educational attainment questions

Recoded category	Old question codes: highest grade attended		New question codes
	Not completed	Completed	
Current Population Survey			
High school dropout	0–12	1–11	31–37
Twelfth grade	...	12	38, 39
Some college	13–16	13–15	40–42
College graduate	17,18	16–18	43–46
Decennial census[1]			
High school dropout	00–14	00–13	01–08
Twelfth grade	...	14	09, 10
Some college	15–18	15–17	11–13
College graduate	19–22	18–22	14–17

[1] Codes for old question in the census are: 00 = never attended school, 01 = nursery school, 02 = kindergarten, 03–14 = 1st through 12th grade, 15–22 = 1 through 8 years of college.

NOTE: Tabulated from a matched sample of individuals 25 to 64 years old from the 1991 and 1992 March Current Population Surveys. Reproduced from D. A. Jaeger, "Reconciling the Old and New Census Bureau Education Questions: Recommendations for Researchers," *Journal of Business and Economic Statistics*, July 1997, pp. 300–09.

Table 3. Cross tabulation of recoded old and new educational attainment questions, by category

Old question education category	New question education category				Row total	Row share	Row match frequency
	Dropout	Twelfth grade	Some college	College graduate			
Dropout	**3,301**	550	34	7	3,892	.145	.848
Twelfth grade	245	**9,216**	718	62	10,241	.383	.900
Some college	23	558	**5,290**	303	6,174	.231	.857
College graduate	6	80	361	**5,997**	6,444	.241	.931
Column total	3,575	10,404	6,403	6,369	26,751	...	...
Column share	.134	.389	.239	.238	...	...	...
Column match frequency	.923	.886	.826	.942	...	...	.890

NOTE: Entries in bold indicate matches between old question and new question. Tabulated from a matched sample of individuals 25 to 64 years old from 1991 and 1992 March Current Population Surveys. Reproduced from D. A. Jaeger, "Reconciling the Old and New Census Bureau Education Questions: Recommendations for Researchers," *Journal of Business and Economic Statistics*, July 1997, pp. 300–09.

tainment questions. The overall rate of matching between the two recoded measures is 89 percent. This varies somewhat by level of educational attainment, with the "college graduate" category having the highest rate on both questions. "Dropout" is matched least frequently in the old question, while "some college" has the lowest rate of matching in the new question. Chart 2 displays the frequency distribution (bars) and cumulative distribution (lines) of educational attainment, using both questions. The distributions are essentially identical, indicating that the observed distribution of educational attainment is unlikely to have been affected by the change in questions, at least when one uses the recoding scheme of table 2. As the number of individuals who are off the diagonal in table 3 makes clear, however, this does not imply that all individuals are in the same category, using both questions. Also, researchers should be aware that the similarity in distribution does not necessarily extend to labor market outcomes such as earnings and unemployment rates.[5]

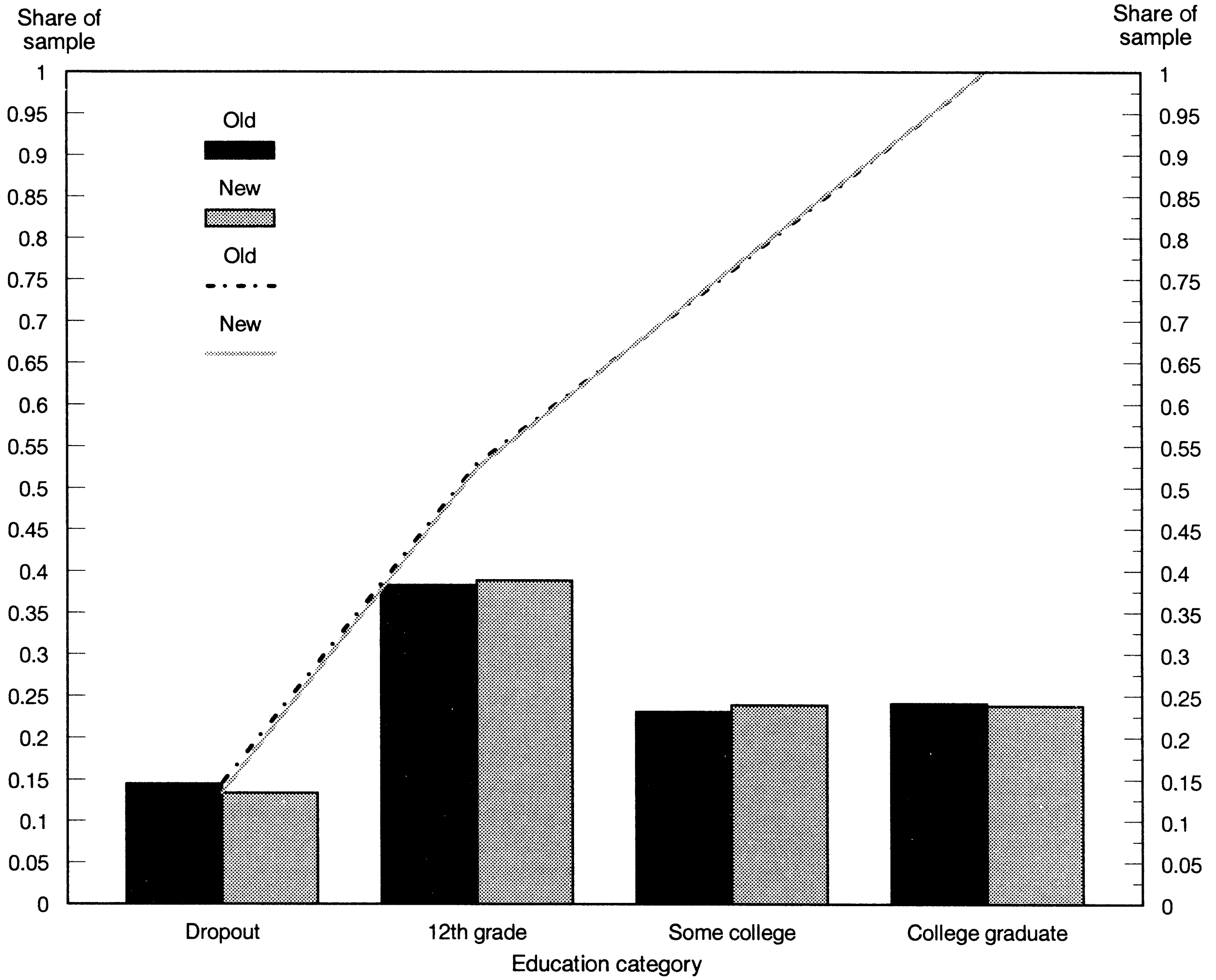
IN SUM, to reconcile the old and new educational attainment questions in the CPS and the decennial census, it is possible to closely approximate "highest grade completed" using the new degree-based question. The preferred method of imputation set forth in this article somewhat overstates the average highest grade completed at the lower end of the educational distribution and somewhat understates the average highest grade completed for individuals who, in fact, attended 2 or 3 years of school after the 12th grade.

It is also possible to create comparable aggregate categories using the old and new educational attainment questions. For the four most widely used categories of educational attainment, the distribution of educational attainment using the recoded old question is essentially identical to that using the recoded new question. □

Footnotes

[1] Robert Kominski and Paul M. Siegel, "Measuring education in the Current Population Survey," *Monthly Labor Review*, September 1993, pp. 34–38.

Chart 2. Distribution and cumulative distribution of recoded old and new education questions, by category



[2] See David A. Jaeger, "Reconciling the Old and New Census Bureau Education Questions: Recommendations for Researchers," *Journal of Business and Economic Statistics*, July 1997, pp. 300–09.

[3] The CPS surveys individuals for 4 months, does not survey them for 8 months, and then again surveys them for 4 months. Because the CPS does not follow individuals who move in between times the survey is administered, it is possible only to match somewhat less than half of a given month's sample to the previous year's survey.

[4] The matching procedure and other sample selection criteria are discussed in greater detail in Jaeger, "Reconciling the Old and New Education Questions."

[5] Jaeger, "Reconciling the Old and New Education Questions"; and H. Frazis and J. Stewart, *Tracking the Returns to Education in the Nineties: Bridging the Gap Between the New and Old CPS Education Items,* Economic Working Paper Number 288 (Bureau of Labor Statistics, September 1996) examine the impact of the change in the question on earnings differentials in detail.