Reconciling the Old and New Census Bureau Education Questions: Recommendations for Researchers

David A. Jaeger
Hunter College and Graduate School, City University of New York, New York, NY 10021, (djaeger@hunter.cuny.edu)

Beginning with the 1990 Census and the January 1992 Current Population Survey (CPS), the Bureau of the Census changed the emphasis of its educational-attainment question from years of education to degree receipt. Using a matched sample from the 1991 and 1992 March CPS, this article addresses how to reconcile the old and new questions. The effects of those methods on the estimated return(s) to education are then examined. Both the estimated linear return and the estimated college–high-school wage differential are slightly larger using information from the new question.

KEY WORDS: Current Population Survey; Return to education; U.S. Census; Wage equations.

Beginning with the 1990 U.S. Census and the January 1992 Current Population Survey (CPS), the Bureau of the Census changed the focus of the educational-attainment question asked in its two major surveys from years of education to degree receipt. The change was motivated by several factors considered important by the Bureau of the Census, including the inaccuracy of inferring degree receipt from the highest grade of completed schooling, the inability to identify specific degrees received, and the growing number of individuals who completed 12th grade without receiving a diploma (Kominski and Siegel 1993).

Researchers who formerly relied on the continuous variable "highest grade completed" must now find new ways to represent educational attainment. In particular, those wishing to use the long time series available in both the decennial Censuses and the CPS require a method to "bridge" the old and new questions. In this article, I explore two methods to create equivalent educational-attainment variables from the old and new questions. First, I examine ways to "linearize" responses to the new (categorical) question to provide a measure that is comparable to "highest grade completed." Second, I propose a categorical recoding scheme for both questions. In both cases I examine how these methods of reconciling responses to the old and new questions affect the estimated return(s) to education. The data used for these analyses are a matched sample from the 1991 and 1992 March CPS that contains responses to both educational-attainment questions as well as demographic and wage information.

Section 1 describes the change in the educational-attainment question and the data used in the article. Section 2 examines methods of linearizing the new question, and Section 3 discusses a consistent categorical recoding scheme. Section 4 concludes. An appendix outlines the procedure used to match consecutive years of the March CPS.

1. DESCRIPTION OF THE EDUCATIONAL–ATTAINMENT QUESTIONS AND THE DATA

Prior to the change in the educational-attainment question, the CPS and Census survey instruments asked individuals what was the highest grade they had attended and whether they had finished that grade. The question consisted of two parts: "What is the highest grade or year of regular school [the individual] has ever attended" and "Did [the individual] complete that grade (year)?" (The second question in the Census instrument is "Did this person finish the highest grade or year attended?") After the change, the question has focused instead on degree receipt: "What is the highest level of school [the individual] has completed or the highest degree [the individual] has received?" The new question gives a choice of categories, which are presented in Table 1. Note that the choices are slightly different in the CPS and the Census; the Census question provides categories for nursery school and kindergarten and combines the "5th or 6th grade" and "7th or 8th grade" categories.

The four–eight–four sampling structure of the CPS permits the matching of individuals across the same survey months of consecutive years. That is, individuals are in the sample for four months, are out for eight, and then are in the sample again for the same four calendar months the following year. Because the educational-attainment question was changed for the 1992 January CPS, individuals in the first four rotations of the 1991 March CPS were asked the old questions in 1991 and the new question in March 1992. I matched individuals from these surveys to create a sample with answers to both questions. Details of the matching procedure are in the Appendix. To reduce the possibility that an individual's true level of schooling would have 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.

Although similar data are also available from the Census Bureau's test of the new question in the 1990 February CPS (see Kominski and Siegel 1993; Park 1994), there are several advantages to using the matched data. First, using data in which the educational-attainment questions were asked

Table 1. Codes for the New Census Bureau Education Questions

CPS code	*Census code*	*Description*
	01	No school completed
	02	Nursery school
	03	Kindergarten
31		Less than 1st grade
32	04	1st, 2nd, 3rd, or 4th grade
33		5th or 6th grade
34		7th or 8th grade
	05	5th, 6th, 7th, or 8th grade
35	06	9th grade
36	07	10th grade
37	08	11th grade
38	09	12th grade, no diploma
39	10	High school graduate—high school diploma, or the equivalent (e.g., GED)
40	11	Some college but no degree
41	12	Associate's degree in college—occupational/ vocational program
42	13	Associate's degree in college—academic program
43	14	Bachelor's degree (e.g., B.A., A.B., B.S.)
44	15	Master's degree (e.g., M.A., M.S., M.Eng., M.Ed., M.S.W., M.B.A.)
45	16	Professional school degree (e.g., M.D., D.D.S., D.V.M., L.L.B., J.D.)
46	17	Doctoral degree (e.g., Ph.D., Ed.D.)

at different points in time should minimize any likelihood that the survey respondents would attempt to reconcile their answers to the two questions. Second, because many researchers use the annual demographic supplement asked in the March CPS, I wanted to gauge the effect of the change in the educational-attainment question on results employing that supplement. In particular, wage data are available for all individuals in my sample. The subsample of the 1990 February CPS for which wage data are available is approximately half the size of my sample. Last, the new question asked in the matched data is the same as that currently in use in the CPS. The test question asked in the 1990 February CPS was similar to that used in the 1990 Census.

A possible disadvantage of using the matched data is that the matching procedure may be nonrandom with respect to characteristics of interest to researchers. Table A.1 in the Appendix shows that there are small differences in the match rates of sex, race, and age subgroups. There are, however, somewhat larger differences between educational categories. In particular, individuals with less than a 7th-grade education are less likely to be matched than those with more education. I found similar results in matching the 1990 and 1991 March CPS and examining match rates by the old educational-attainment question as well as by the new educational-attainment question in matching the 1991 and 1992 March CPS, indicating that this phenomenon is not an artifact of the change in the question. Given the small number of individuals with this level of educational attainment (2.3%), it is unlikely that large biases are introduced into my results. In addition, as will be noted, I replicated all of the analysis in the article using the February 1990 CPS and found quite similar results.

2. LINEARIZATION OF THE NEW EDUCATIONAL–ATTAINMENT QUESTION

This section explores three different ways to create a linear variable representing "highest grade completed" from the categorical responses to the new educational-attainment question. Because much of the existing literature on the returns to education uses a linear term to capture differences in educational attainment, these methods permit the estimation of a comparable return after the change in the educational-attainment question. A natural choice for imputing values for highest grade completed from the new question is to use some measure of central tendency, conditional on each value of the new question. Park (1994) suggested using the mean observed highest completed grade for each level of the new question. Because the mean is sensitive to outliers, other measures of central tendency may perform better, however. Medians are less sensitive to outliers than means, but modal values provide the greatest number of observations whose imputed value is the same as the observed value. I therefore also examine using conditional median and modal values to impute a value for highest grade completed as well as a slight modification of the recoding implied by the median values.

Table 2, columns 4 and 6, presents the mean and median/modal values of highest grade completed for each level of the new educational-attainment question. Note that the median and modal values are the same for each level of the new question; I will subsequently refer to this imputation method as that based on medians. One measure of the accuracy of these imputations is how well they predict, on average, the actual highest grade completed. Table 3 presents the mean imputed highest grade completed for each of the 19 values of highest grade completed, and Figure 1 presents these means graphically. A similar table and figure using the 1990 February CPS is available from the author by request. The 45-degree line in Figure 1 indicates agreement between the mean imputed values and the observed values.

The average imputed values are substantially higher than the observed values for individuals with an 8th-grade education or less, with the imputations using medians performing better than those using means. To address this overestimation, I modified the median-based imputation scheme by assigning individuals in categories "1st through 4th grade," "5th or 6th grade," and "7th or 8th grade," the midpoint of the range represented by the category. That is, I assigned the values 2.5, 5.5, and 7.5 to categories "1st through 4th grade," "5th or 6th grade," and "7th or 8th grade," respectively. The "assigned" values are also listed in column 7 of Table 2. The average imputations using these "assigned" values are closer to the 45-degree line than the median imputations but still substantially overrepresent the educational attainment of individuals with two or fewer years of education. The imputed values also underrepresent the highest grade completed for individuals with 14 or more years of education, with the largest difference among individuals with 15 years of education. Among those who reported 15 years under the new question, nearly all are given imputed

Table 2. Imputations of Highest Grade Completed (HGC) for New Education Question

			Observed HGC			
New question category	CPS code	N	Mean	Std. dev. of mean	Median/ mode	Assigned HGC
Less than 1st grade	31	91	1.30	3.17	0	0.
1st, 2nd, 3rd, or 4th grade	32	217	3.92	2.67	3	2.5
5th or 6th grade	33	339	6.22	2.23	6	5.5
7th or 8th grade	34	814	7.84	1.44	8	7.5
9th grade	35	543	9.08	1.33	9	9.
10th grade	36	836	9.90	1.02	10	10.
11th grade	37	735	10.81	0.86	11	11.
12th grade, no diploma	38	309	11.38	1.64	12	12.
H.S. graduate or equivalent	39	10,095	12.00	0.83	12	12.
Some college, no degree	40	4,620	13.35	1.18	13	13.
Associate's degree—occ./voc.	41	923	13.87	1.17	14	14.
Associate's degree—academic	42	860	14.29	1.04	14	14.
Bachelor's degree	43	4,174	16.04	0.90	16	16.
Master's degree	44	1,623	17.57	0.99	18	18.
Professional school degree	45	345	17.71	0.92	18	18.
Doctoral degree	46	227	17.84	0.65	18	18.

NOTE: Tabulated from a matched sample of individuals 25 to 64 years old from the 1991 and 1992 March CPS. The median and modal values of highest grade completed were the same for all categories of the new question.

values between 13 and 14.29 years, depending on the imputation method and the category they reported under the new question.

One of the principal uses of the educational-attainment question in the CPS is to estimate the return(s) to education in wage equations. Researchers who want to examine changes over time in those returns may well be concerned that the change in the educational-attainment question will cause spurious changes in the estimated returns. To gauge the magnitude of this artifact, I estimated the conventional log wage equation using the observed highest grade completed and the three different imputation methods discussed previously. For the regression sample, individuals who worked on farms; who had nonpositive or allocated weeks worked, usual hours worked, or annual earnings; or whose hourly wage was less than $1 or greater than $200 were excluded; log wages are defined at log[annual earnings/(weeks worked × usual hours worked)]. Table 4 presents results that control for potential labor-market experience in two different ways. Columns 3–6 use the usual potential experience variable, defined as (age − highest grade completed − 6), where education is the measure of education (actual or imputed) used in the regression. Because potential experience is created using years of education, any mismeasurement that the imputation process introduces will also contaminate this variable. In columns 7–10, I therefore also present models that use age in place of potential experience. Additional covariates are dummy variables for "nonwhite" and "female." I report heteroscedasticity-consistent jackknife (Efron 1982) standard errors; conventional ordi-

Table 3. Mean Imputed Highest Grade Completed by Actual Highest Grade Completed for Different Imputation Methods

		Imputation method		
Actual	N	Means	Medians	Assigned
0	95	2.56	1.48	1.38
1	15	5.73	5.00	4.63
2	63	5.55	4.84	4.50
3	73	4.72	3.96	3.50
4	76	5.58	5.01	4.59
5	105	6.40	6.13	5.77
6	264	7.02	6.84	6.41
7	240	7.87	7.98	7.53
8	609	8.25	8.34	7.96
9	656	9.47	9.47	9.43
10	888	10.26	10.36	10.35
11	808	11.12	11.29	11.29
12	10,958	12.12	12.10	12.09
13	1,930	13.28	13.02	13.02
14	2,677	13.67	13.46	13.46
15	850	13.95	13.74	13.74
16	3,759	15.86	15.82	15.82
17	662	16.58	16.73	16.73
18	2,023	17.28	17.59	17.59

NOTE: Tabulated from a matched sample of individuals 25 to 64 years old from the 1991 and 1992 March CPS.

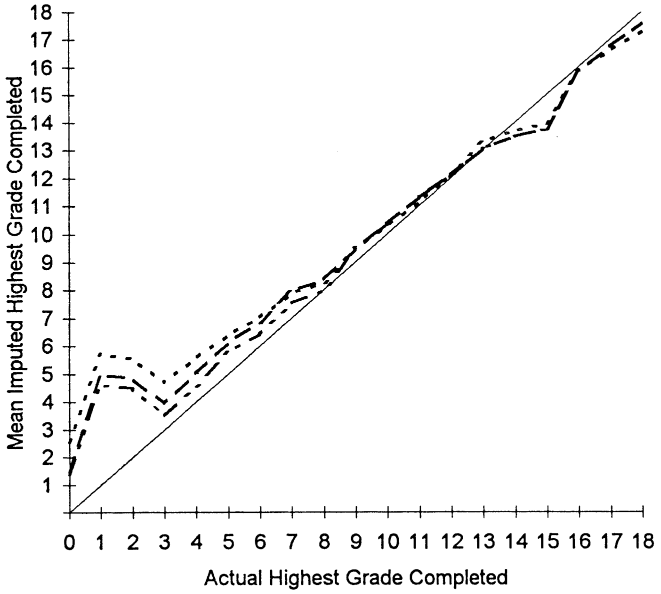

Figure 1. Mean Imputed Highest Grade Completed Versus Actual Highest Grade Completed: Imputation Method: - - -, Means; ———, Medians; — - - —, Assigned.

Table 4. Estimated Return to Education Using Actual and Imputed Highest Grade Completed (jackknife standard errors of coefficients in parentheses, p values of χ^2 statistics in brackets)

		Potential experience				Age			
			Imputation method				Imputation method		
Subgroup	N	Actual	Mean	Median	Assigned	Actual	Mean	Median	Assigned
Full sample	19,230	.093	.102	.098	.097	.085	.094	.091	.089
		(.002)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)
χ^2_1			155.660	56.488	27.285		152.584	56.036	26.712
			[.000]	[.000]	[.000]		[.000]	[.000]	[.000]
Sex									
Men	9,959	.086	.095	.092	.090	.076	.085	.082	.080
		(.002)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)
χ^2_1			119.198	46.489	26.226		117.718	56.529	26.094
			[.000]	[.000]	[.000]		[.000]	[.000]	[.000]
Women	9,271	.102	.111	.107	.105	.097	.105	.102	.100
		(.003)	(.003)	(.003)	(.003)	(.002)	(.003)	(.003)	(.002)
χ^2_1			44.945	15.492	6.230		42.542	14.571	5.676
			[.000]	[.000]	[.013]		[.000]	[.000]	[.017]
Race									
Whites	16,946	.092	.100	.097	.095	.084	.093	.089	.088
		(.002)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)
χ^2_1			130.028	44.479	20.861		125.024	42.670	19.621
			[.000]	[.000]	[.000]		[.000]	[.000]	[.000]
Nonwhites	2,284	.101	.113	.110	.108	.092	.104	.101	.099
		(.005)	(.006)	(.006)	(.005)	(.005)	(.005)	(.005)	(.005)
χ^2_1			25.883	13.392	7.697		29.121	15.632	8.844
			[.000]	[.000]	[.006]		[.000]	[.000]	[.003]
Age									
25–35	5,401	.109	.116	.113	.112	.085	.092	.090	.088
		(.004)	(.004)	(.004)	(.004)	(.003)	(.003)	(.003)	(.003)
χ^2_1			20.201	8.802	4.858		18.812	7.244	3.026
			[.000]	[.003]	[.028]		[.000]	[.007]	[.082]
36–46	6,651	.095	.105	.100	.099	.086	.096	.092	.091
		(.004)	(.004)	(.004)	(.004)	(.003)	(.003)	(.003)	(.003)
χ^2_1			63.283	17.842	13.545		69.465	23.510	15.132
			[.000]	[.000]	[.000]		[.000]	[.000]	[.000]
47–64	7,178	.081	.090	.087	.085	.084	.093	.089	.087
		(.003)	(.003)	(.003)	(.003)	(.002)	(.003)	(.003)	(.003)
χ^2_1			61.633	22.822	10.061		64.279	23.943	9.225
			[.000]	[.000]	[.002]		[.000]	[.000]	[.002]

NOTE: Estimated by OLS. Dependent variable is log(hourly wage). Potential experience columns include potential experience and (potential experience)^2/100 as covariates. Age columns include (age − 25) and (age − 25)^2/100 as covariates. Where appropriate, models also include dummy variables for female and nonwhite. χ^2 is from a Wald test for equality of return using actual and imputed highest grade completed. Standard errors and covariances between actual and imputed estimates are computed using the jackknife (Efron 1982). Data are individuals 25 to 64 years old from a matched sample of the 1991 and 1992 March CPS.

nary least squares (OLS) standard errors, as well as those estimated using White's (1980) method, are essentially the same in the number of decimal places reported.

Table 4 also presents Wald tests of the equality of the return to education using actual and imputed highest grade completed. The Wald statistic, w, is

$$w = (R\hat{\beta})'(R\hat{V}R')^{-1}(R\hat{\beta}) \sim \chi^2_p, \qquad (1)$$

where $\hat{\beta}$ is the vector of the old and new estimates of the return to education, $\hat{V}$ is an estimate of the variance–covariance matrix for those estimates, R is a set of linear restrictions on the coefficients such that $R\hat{\beta} = 0$, and p is the number of rows in R. In this case $R = [1 - 1]$ and $p = 1$. The conventional OLS estimator of V assumes that the error variance in both the old and new regressions is homoscedastic. It is possible that misspecification of the education variable in either or both regressions will induce heteroscedasticity in the error terms, however. If this is the case, tests based on the OLS estimator of V will have the wrong size. In addition, because the returns to education using both the old and new variables are estimated with the same data, the covariance between the estimates is likely to be substantial. To address both of these issues, I use the jackknife (Efron 1982) to estimate V.

For the full sample (row 1 of Table 4), using either potential experience or age to control for labor-market experience, the estimated return to education is somewhat higher using the imputed highest completed grade rather than the observed values. This is due in large part to the underrepresentation (on average) of years of schooling at the upper end of the educational distribution in the imputation schemes. Of the three imputation methods, that using assigned values for the eighth grade and below performs best and is a slight improvement over using median values, indicating that a small part of the overestimation of the return to education also derives from the overrepresentation of years of educational attainment among individuals at the lower end of the educational distribution. Results using 1991 wage data as well as data from the 1990 February CPS give the same qualitative conclusions and are available from the author by

Table 5. Cross-tabulation of Old (highest grade completed) by New Education Variables

Old	New question category																	
question (HGC)	<1st	1st–4th	5th–6th	7th–8th	9th	10th	11th	12th, no dip.	H.S. grad.	Some coll.	Occ. assoc.	Acad. assoc.	Bach.	Mast.	Prof.	Doct.	Row total	Row freq.
0	72	10	5	4	0	0	0	2	1	1	0	0	0	0	0	0	95	.004
1	1	11	0	0	0	0	0	0	2	0	0	0	0	1	0	0	15	.001
2	6	39	3	1	2	0	0	0	12	0	0	0	0	0	0	0	63	.002
3	1	60	3	4	0	1	1	1	1	0	1	0	0	0	0	0	73	.003
4	0	44	16	5	6	1	0	2	1	1	0	0	0	0	0	0	76	.003
5	1	10	81	5	3	0	0	0	4	1	0	0	0	0	0	0	105	.004
6	1	17	167	45	2	7	3	1	17	3	0	0	0	1	0	0	264	.010
7	2	3	16	197	7	4	2	2	7	0	0	0	0	0	0	0	240	.009
8	1	7	12	449	83	21	3	4	24	3	0	0	2	0	0	0	609	.023
9	1	5	8	47	338	173	16	8	54	4	1	0	1	0	0	0	656	.025
10	0	2	3	15	49	521	174	17	102	5	0	0	0	0	0	0	888	.033
11	0	1	0	5	15	42	441	83	205	12	2	0	1	1	0	0	808	.030
12	5	6	21	32	35	62	87	173	9,186	1,121	123	36	60	7	3	1	10,958	.410
13	0	0	0	3	2	1	5	8	217	1,486	111	57	35	4	1	0	1,930	.072
14	0	1	1	1	0	1	3	4	155	1,344	519	538	94	11	4	1	2,677	.100
15	0	1	0	1	0	0	0	1	30	463	90	120	129	6	7	2	850	.032
16	0	0	3	0	0	2	0	3	54	142	63	83	3,251	141	12	5	3,759	.141
17	0	0	0	0	0	0	0	0	6	14	6	9	338	263	16	10	662	.025
18	0	0	0	0	1	0	0	0	17	20	7	17	263	1,188	302	208	2,023	.076
Column total	91	217	339	814	543	836	735	309	10,095	4,620	923	860	4,174	1,623	345	227	26,751	
Column freq.	.003	.008	.013	.030	.020	.031	.027	.012	.377	.173	.035	.032	.156	.061	.013	.008		

NOTE: Tabulated from a matched sample of individuals 25 to 64 years old from the 1991 and 1992 March CPS.

request. Following Card and Krueger's (1992) suggestion, I also estimated these models after recoding all individuals who completed 7th grade or less (or had an imputed value indicating 7th grade or less) to having completed 7th grade. The results were similar to those in the first row of Table 4, reinforcing the conclusion that much of the difference in returns between the imputation methods and the actual values derives from differences at the upper end of the educational distribution.

The remainder of Table 4 presents results from the estimation of these models for sex, race, and age subgroups. Covariates include, where appropriate, dummy variables for nonwhite and female. The imputations used are taken from Table 3. That is, I did not separately compute means and medians for each subgroup. For all subgroups the results echo those for the full sample. The "assigned" imputation method gives results that are closest to, although always larger than, those estimated using observed values for highest grade completed.

In summary, the estimated linear return to education is somewhat higher using the various imputation methods, largely due to underrepresentation of highest grade completed for individuals who, in reality, attended 14 or 15 years of school. Although for most of the subgroups I can reject the hypothesis that the return estimated with the actual highest grade completed is equal to the return estimated with any of the imputation methods, the "assigned" method produces an estimated return that is closest to that estimated using actual responses on highest grade completed, and is between .3 and .7 percentage points higher, depending on the subgroup. This method would appear to provide the greatest comparability in estimating a linear return to education across the break in the educational-attainment question.

3. A CONSISTENT CATEGORICAL RECODING SCHEME

A second strategy in reconciling the old and new educational-attainment questions is to create categorical variables from the responses to the old and new questions—that is, recode the old question to be comparable to the new question. This categorization should be useful to researchers who want to estimate the returns to specific categories of educational attainment (e.g., the college–high-school wage premium) or for researchers who want to stratify their samples by education categories. Siegel (1991)

Table 6. Categorical Recoding Scheme for Old and New Education Questions

	Old question: highest grade attended		
Recoded category	Not completed	Completed	New question code
	Current Population Survey		
Dropouts	0–12	1–11	31–37
12th grade		12	38, 39
Some college	13–16	13–15	40–42
College graduates	17, 18	16–18	43–46
	Census		
Dropouts	00–14	01–13	01–08
12th grade		14	09, 10
Some college	15–18	15–17	11–13
College graduates	19–22	18–22	14–17

NOTE: The codes for the old question in the Census are 00 = never attended school, 01 = nursery school, 02 = kindergarten, 03–14 = 1st through 12th grades, 15–22 = 1 through 8 years of college. The codes for the new question are presented in Table 1.

Table 7. Cross-tabulation of Categorically Recoded Old and New Variables

	New question						
Old question	Dropouts	12th grade	Some college	College grads.	Row total	Row share	Row match freq.
Dropouts	**3,301**	550	34	7	3,892	.145	.848
12th grade	245	**9,216**	718	62	10,241	.383	.900
Some college	23	558	**5,290**	303	6,174	.231	.857
College graduates	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 freq.	.923	.886	.826	.942			.890

NOTE: Entries indicated matches are shown in bold. Tabulated from a matched sample of individuals 25 to 64 years old from the 1991 and 1992 March CPS.

presented extensive evidence that the overall distribution of education attainment is roughly equivalent with either question. A cross-tabulation of the (not recoded) old and new questions is presented in Table 5, and is similar to those tabulated from the 1990 February CPS by Kominski and Siegel (1993) and Park (1994). In creating a consistent categorical recoding scheme to reconcile the new and old questions, I attempted to preserve this similarity in the distribution to the greatest extent possible.

Although it is possible to preserve much of the information in the new question, most researchers are interested in four educational subgroups—high-school dropouts and below, high-school graduates (or, more accurately, individuals who have completed 12th grade), individuals with some college, and college graduates. Note that the old question did not provide any information about high-school diploma receipt; I will use the rubric "12th grade" rather than "high-school graduates" to reflect this fact. My recoding scheme attempts to provide consistent codings for these categories. I experimented with several schemes that both did and did not take into account information from the old question about whether the highest grade was completed. The scheme that provided the best conceptual agreement between the two questions and a high match rate between the recoded variables is presented in Table 6. The top panel presents the recoding scheme for the CPS, and the bottom panel contains the recoding scheme for the Census.

The recoding scheme for the old question uses, for the most part, highest grade completed. The old question is recoded into the category "Dropouts" in a straightforward way. This category includes all individuals who attended 11th grade or less and individuals who attended, but did not complete, 12th grade. The "12th grade" category comprises all individuals who completed 12th grade. Individuals who reported attending 13, 14, or 15 years of school (regardless of completion status), as well as individuals who attended but did not complete 16 years, are recoded into the "Some college" category. It is important to note that individuals who attended, but did not complete, a 13th year of school are counted as having some college because these individuals are often included among individuals with 12 years of school rather than as college attenders. The match between the old and new recoded variables is improved by including these individuals among individuals with some college experience, however, because 74.5% of them report "Some college but no degree" on the new question. Individuals who completed 16 or more years of education are recoded into the "College graduate" category.

Although this recoding scheme groups all college graduates, some researchers may be interested in dividing this last category further (i.e., into "College" and "Post-college" categories). I recommend that individuals who reported finishing 16 or 17 years of education be recoded into a "4 or 5 years college" category. Although this recoding does not correspond to the usual practice of imputing a college diploma from 16 completed years of education, the match is improved by including individuals who completed 17 years with those who completed 16 years. Among individuals who reported attending 17 years of school with the old question, 57.9% reported receiving only a bachelor's degree with the new question, and 35.3% reported receiving a master's degree or higher.

The new question is recoded as one would expect into "Dropout." The categories "12th grade, no diploma" and "High school graduate" are both recoded into "12th grade." Individuals who reported completing 12th grade without receiving a diploma with the new question were more likely than not (55.3%) to report finishing 12th grade with the old question. Individuals who reported receiving an associate's degree of either type are recoded into "Some college" as are those who reported attending college without receiving a degree. Those who received a bachelor's, master's, professional, or doctoral degree are recoded into the "College graduate" category.

A cross-tabulation of the recoded old and new education variables is presented in Table 7; a version of this table using the 1990 February CPS is available from the author by request. The match rate between the recoded new and old categories is 89.0% for the full sample and varies little across age and sex subgroups. Nonwhites are slightly less likely to fall in the same category (85.3%) than whites (89.5%).

In terms of the returns to education, Table 8 presents estimates from a model that includes dummy variables for three of the educational-attainment categories ("12th grade" is the reference category), (age − 25), (age − 25)^2/100, and (where appropriate) dummy variables for "nonwhite" and "female," estimated with the recoded old and new educational-attainment variables. As in Table 4, I report standard errors calculated using the jackknife (Efron 1982). I also report Wald tests for the null hypothesis that the indi-

Table 8. Estimated Returns to Education Using Categorical Recoding (jackknife standard errors of coefficients in parentheses, p values of χ^2 tests in brackets)

			Sex				*Race*				*Age*					
Education	*Full sample*		*Men*		*Women*		*Whites*		*Nonwhites*		*25–34*		*35–44*		*45–64*	
category	*Old*	*New*	*Old*	*New*	*Old*	*New*	*Old*	*New*	*Old*	*New*	*Old*	*New*	*Old*	*New*	*Old*	*New*
Coefficients																
Dropouts	-.261	-.261	-.251	-.253	-.294	-.288	-.258	-.259	-.274	-.264	-.246	-.247	-.240	-.241	-.286	-.280
	(.014)	(.014)	(.019)	(.020)	(.020)	(.020)	(.015)	(.016)	(.035)	(.036)	(.026)	(.027)	(.025)	(.027)	(.021)	(.021)
12th grade	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
Some college	.154	.170	.117	.135	.184	.198	.155	.170	.151	.172	.137	.160	.174	.179	.143	.165
	(.010)	(.010)	(.014)	(.014)	(.014)	(.014)	(.011)	(.011)	(.029)	(.028)	(.017)	(.017)	(.017)	(.017)	(.018)	(.002)
Coll. grads.	.450	.465	.398	.419	.495	.505	.445	.456	.491	.542	.421	.430	.460	.476	.452	.469
	(.010)	(.010)	(.014)	(.014)	(.014)	(.015)	(.011)	(.011)	(.031)	(.031)	(.018)	(.018)	(.017)	(.017)	(.017)	(.017)
Intercept	2.206	2.195	2.156	2.149	1.893	1.878	2.217	2.207	2.037	2.018	2.098	2.091	2.352	2.331	2.344	2.316
	(.013)	(.013)	(.017)	(.017)	(.017)	(.017)	(.013)	(.013)	(.038)	(.037)	(.023)	(.023)	(.187)	(.187)	(.195)	(.195)
Hypothesis tests (old = new)																
Dropouts	.001		.061		.220		.017		.203		.004		.005		.250	
	[.979]		[.805]		[.639]		[.896]		[.652]		[.949]		[.943]		[.617]	
Some college	5.551		3.890		2.290		4.487		1.102		3.796		.210		3.724	
	[.018]		[.049]		[.130]		[.034]		[.294]		[.051]		[.647]		[.054]	
Coll. grads.	12.786		12.913		2.332		6.703		9.120		1.490		5.089		5.509	
	[.003]		[.000]		[.127]		[.010]		[.003]		[.222]		[.024]		[.019]	
Intercept	11.351		3.726		8.721		7.658		3.075		1.278		.550		.863	
	[.001]		[.054]		[.003]		[.006]		[.079]		[.258]		[.458]		[.353]	
All old = all new	24.397		21.457		11.475		15.871		12.383		5.600		7.683		8.956	
	[.000]		[.000]		[.022]		[.003]		[.015]		[.231]		[.104]		[.062]	
N	19,230		9,959		9,271		16,946		2,284		5,401		6,651		7,178	

NOTE: Estimated by OLS. Dependent variable is log(hourly wage). All models include (age — 25) and (age — $25)^2/100$ as covariates. Where appropriate, models also include dummy variables for female and nonwhite. χ^2 is from a Wald test of equality of old and new education variables and intercepts. Standard errors and covariances between estimates calculated using the jackknife (Efron 1982). Data are individuals 25 to 64 years old from a matched sample of the 1991 and 1992 March CPS.

vidual education coefficients are equal as well as for the null hypothesis that the three education coefficients and the constant term are jointly equal across the old and new variables. These are calculated using Equation (1) and the jackknife to estimate V with the appropriate matrices for R. Versions of Table 8 using the 1990 February CPS and the matched sample with 1991 wage data are available from the author by request.

Columns 2 and 3 of Table 8 present results for the full sample. I can reject the joint hypothesis that the coefficients on the old and new variables are equal, although the results are generally comparable. Differences in the point estimates may be quantitatively important for examining changes in the returns to education, however, especially over short periods of time. Of particular interest is the college–high-school wage premium, which is estimated to be 56.8% with the recoded old variables and 59.2% with the recoded new variables. This difference of 2.4 percentage points is relatively small compared to the growth in the differential of 11.7% between 1979 and 1987 or 9% between 1963 and 1987 (Katz and Murphy 1992). Comparisons that span the change in the educational-attainment question made over shorter periods of time, when changes in the differential would likely be smaller, may suffer from larger proportional biases.

This difference should be considered something of an upper bound on the magnitude of the change in the college–high-school wage differential due to the change in the educational-attainment question. Results from the 1990 February CPS give "College graduate" coefficient estimates from the old and new recoded questions of .464 and .474, respectively. That is, the difference in the estimated premia is 1.6 percentage points. Moreover, results from the matched sample using 1991 wages shows *no* difference in the estimated college premium, with both old and new "College graduate" coefficients being estimated as .460. This discrepancy between the estimated returns using 1991 wages and those using 1992 wages is the subject of ongoing research at the Bureau of Labor Statistics.

The remaining columns of Table 8 present results for the same models estimated with sex, race, and age subgroups. Figure 2 shows the total (lines) and marginal (bars) returns to education based on these models for men and women; graphs for the other subgroups are similar. The estimated returns to college attendance and college graduation are consistently somewhat larger with the new variables than with the old. As an example of the possible effects of the change in the educational-attainment question on comparisons over short periods of time in the returns to education, Bound and Johnson (1995), using CPS data, estimated that between 1988 and 1993 the annual growth in the college–high-school wage differential was .7% for men and .9% for women. The results for men and women in columns 4–7 of Table 8 indicate that the change in the educational-attainment question may have contributed 3.1 and 1.6 percentage points to estimated growth in that differential for men and women, respectively. (These differences are 1.8 and 1.5 percentage points, respectively, in the

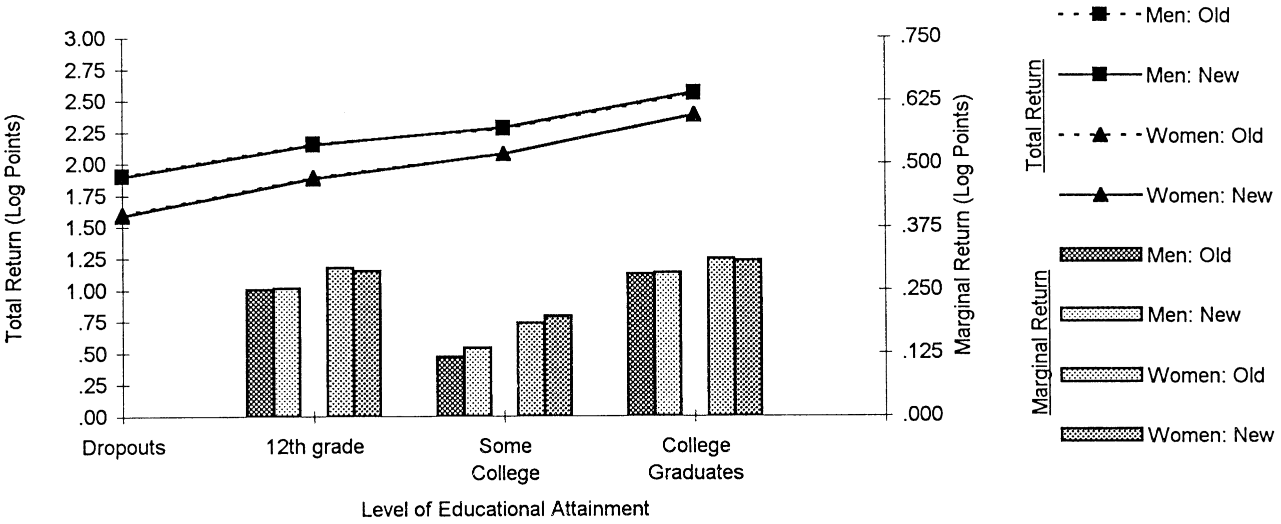

Figure 2. Total and Marginal Returns to Education Using Recoded Old and New Variables: Men and Women.

1990 February CPS, but using 1991 wages with the matched sample gives differences of .3 and −.3 percentage points, respectively.) That is, the total change in the differential between 1988 and 1993 may have been .4% (as opposed to 3.5%) for men and 2.9% (as opposed to 4.5%) for women. Although Bound and Johnson's (1995) conclusions are not qualitatively altered by these differences (and are actually strengthened), the quantitative effects are somewhat substantial.

4. CONCLUSION

Although the change in the educational-attainment question in the U.S. Census and the CPS creates some difficulties for researchers wishing to use the long time series available with those data, I showed that it is possible to reconcile quite closely the old and new questions. I proposed several different methods for imputing highest grade completed to responses to the new question. Using a matched sample from the 1991 and 1992 March CPS, I found that the method using median values for highest grade completed for all but the lowest educational-attainment categories gave results that were closest to those estimated using actual highest grade completed. The difference between the estimated return using actual and imputed values with this method was between .3 and .7 percentage points, depending on the subsample.

I also proposed a categorical recoding scheme for researchers who want to stratify their samples by educational attainment or who want to estimate more flexible functions for the returns to education. This categorical recoding scheme provides an 89.0% match rate between the recoded responses to the old and the new questions and provided nearly identical distributions of educational attainment in my sample. In addition, I found that the estimated returns to education were generally comparable with this recoding scheme, with the return to college attendance and college graduation (relative to 12th grade) being somewhat larger with the new question. These differences in the point estimates may be quantitatively important for comparisons over time. Of particular interest to many researchers is the college–high-school wage premium. For the full sample, I found that the estimated premium was 2.4 percentage points higher using the recoded new question and varied somewhat between sex, race, and age subgroups.

Although these differences may not have large quantitative or qualitative effects on comparisons over long time periods, researchers examining changes over shorter periods that include the break in the educational-attainment question should be cautious, particularly when looking at subsamples. Some of the observed changes in the returns to education may be an artifact of the change in the educational-attainment question and not a reflection of changing economic realities. The results in this article provide a benchmark for the magnitude of the effects of the change in the educational-attainment question.

ACKNOWLEDGMENTS

All work on this article was completed while I was a Research Economist at the U.S. Bureau of Labor Statistics. I thank John Bound for the original impetus for this article as well as for comments on early versions. I also thank Charles Brown, Harley Frazis, Marilyn Manser, Marianne Page, Jay Stewart, Sarah Turner, and two anonymous referees for helpful comments on earlier versions and Lisa Neidert for assistance with data and obtaining Bureau of the Census documents. A version of this article entitled "The New Current Population Survey Education Question: A Recommendation" was presented at the Population Association of America meetings in May 1994. The views expressed in this article are solely those of the author and do not necessarily reflect the policies of the Bureau of Labor Statistics.

APPENDIX: MATCHING 1991 AND 1992 MARCH CURRENT POPULATION SURVEYS

The procedure I used to match consecutive years of the March CPS is similar to (although more conservative than) the procedure outlined by Welch (1993). The steps of the matching procedure were as follows:

Table A.1. Match Rates by Sample Characteristics

Subgroup	*Matched sample*	*Matched sample/ potential matches*	*Regression sample/ matched sample*
Full sample	26,751	.925	.716
Sex			
Men	12,698	.922	.778
Women	14,053	.927	.660
Race			
White	23,484	.928	.722
Nonwhite	3,267	.898	.699
Age			
25–34	6,907	.891	.782
35–44	8,519	.926	.781
45–64	11,325	.945	.634
Educational attainment (new question)			
Less than 1st grade	91	.722	.330
1st, 2nd, 3rd, or 4th grade	217	.783	.419
5th or 6th grade	339	.731	.478
7th or 8th grade	814	.883	.467
9th grade	543	.896	.501
10th grade	836	.911	.578
11th grade	735	.907	.623
12th grade, no diploma	309	.883	.583
H.S. graduate or equivalent	10,095	.932	.699
Some college, no degree	4,620	.933	.749
Associate's degree—occ./voc.	923	.926	.802
Associate's degree—academic	860	.937	.799
Bachelor's degree	4,174	.939	.813
Master's degree	1,623	.951	.859
Professional school degree	345	.938	.719
Doctoral degree	227	.954	.837

NOTE: Match rates are from matching the 1991 and 1992 March CPS. "Matched sample" is the number of individuals used in the analysis. "Potential matches" are the number of individuals in the 1992 file after Step 3. "Regression sample" is individuals used in estimating regressions. See text for further description.

1. *Make original extraction:* All individuals 23–64 years old in 1991 (24,115 households with 42,037 individuals) and all individuals 25–64 years old in 1992 (23,623 households with 40,002 individuals) were extracted.

2. *Delete allocations:* Individuals whose line number, sex, race, highest grade completed, or enrollment status were allocated were deleted in both years. Individuals who were enrolled in school in either year were deleted. In addition, individuals who in 1992 reported not living in the same house the previous year were deleted. This left 23,849 households with 40,947 individuals in 1991 and 19,635 households with 32,970 individuals in 1992.

3. *Keep matching households:* Only the 17,151 households whose household identification number appeared in both years were kept, leaving 30,191 individuals in the 1991 file and 28,932 individuals in the 1992 file.

4. *Keep matching individuals:* Both years were sorted and matched by household identification number and the individual's line number within the households. Individuals who appear in the same household on the same line number are considered matched if their reported sex and race are the same in both years and their reported age is 0, 1, or 2 years greater in 1992 than in 1991. This gives a matched sample of 26,751 individuals from 16,298 households. Note that of the 28,932 maximum potential matches (the number of individuals in the 1992 file after Step 3), 92.5% were matched. More than 95.0% of the households whose identification numbers were common to both years are represented by these individuals. Table A.1 presents match rates for various subgroups of the sample.

Welch's (1993) procedure then attempts to match the remaining unmatched records; my procedure does not.

The samples used for the regression analysis exclude individuals living on farms or whose wage and salary income, weeks worked, or usual hours worked were allocated or nonpositive, or whose hourly wage was less than $1 or greater than $200.

[Received July 1994. Revised May 1996.]

REFERENCES

Bound, J., and Johnson, G. (1995), "What Are the Causes of Rising Wage Inequality in the United States?" *Federal Reserve Bank of New York Economic Policy Review*, 1, 9–17.

Card, D., and Krueger, A. B. (1992), "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy*, 100, 1–40.

Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia: Society for Industrial and Applied Mathematics.

Katz, L., and Murphy, K. M. (1992), "Changes in Relative Wages, 1963–1987: Supply and Demand Factors," *Quarterly Journal of Economics*, 107, 35–78.

Kominski, R., and Siegel, P. M. (1993), "Measuring Educational Attainment in the Current Population Survey," *Monthly Labor Review*, 116(9), 34–38.

Park, J. H. (1994), "Estimation of Sheepskin Effects and Returns to Schooling Using the Old and the New CPS Measures of Educational Attainment," unpublished draft, Princeton University, Dept. of Economics.

Siegel, P. M. (1991), "Note on the Proposed Change in the Measurement of Educational Attainment in the Current Population Survey," unpublished draft, U.S. Bureau of the Census.

Welch, F. (1993), "Matching the Current Population Surveys," *Stata Technical Bulletin*, 12, 7–9.

White, H. (1980), "Heteroskedasticity-Consistent Covariance Matrix Estimation and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838.