Comparing earnings inequality using two major surveys

Some previous research suggests that discrepancies exist between the National Longitudinal Survey of Youth and the Current Population Survey in terms of earnings trends; when the sample is limited to full-time, year-round workers, however, the discrepancies are largely eliminated

Mark S. Handcock,
Martina Morris,
and
Annette Bernhardt

Much of the research on the growing dispersion of earnings has relied on the March supplement to the Current Population Survey (CPS). As the research questions have turned to such issues as job instability and long-term wage growth, however, the focus often has shifted to longitudinal surveys, such as the Panel Study of Income Dynamics (PSID)[1] and the National Longitudinal Surveys (NLS).[2] In a recent unpublished but widely cited paper,[3] Peter Gottschalk and Robert A. Moffitt compare annual earnings trends from the PSID and two cohorts of the NLS with those of the CPS.[4] The authors find that reported earnings in the PSID and the original NLS cohort show roughly the same trends as the CPS, although the magnitudes are quite different.

For the later NLS cohort, however, known as the National Longitudinal Survey of Youth 1979 (NLSY79), Gottschalk and Moffitt find both significantly lower variance in reported annual earnings and a negative trend in variance over time (1979–88)—at least for high school graduates. In addition, a more recently published paper using different methodology finds a similar discrepancy.[5] Because the findings of these studies stand in sharp contrast to the well-known "stylized fact" that the variance in earnings was increasing substantially during the 1980s, serious questions may be raised about the validity of the NLSY79 for research on the topic of recent trends in earnings inequality.

This article focuses on the comparison between the NLSY79 and the CPS, updating the Gottschalk-Moffitt analysis to 1994, the final year of data collection for the NLSY79 cohort. Because Gottschalk and Moffitt report few discrepancies in the trends for high school dropouts, the analysis is restricted to high school graduates. The article begins by replicating the Gottschalk-Moffitt analysis in order to verify the discrepancies in reported earnings between the two sets of data. Next, exploratory data analysis and respecified regression models are used to compare the trends and patterns, and to look for potential sources of the discrepancies. The final section discusses the implications of the findings for the validity of the two samples.

Data and methods

The present study generally follows the conventions adopted by Gottschalk and Moffitt. For their benchmark analyses, they select white males in the civilian noninstitutionalized population and divide the samples into cells defined by single years of age (from 16 to 31 years), level of education (less than a high school education, high school graduate or more), and survey year (1979-88).[6] Nominal annual earnings are adjusted for inflation and are expressed in constant (1982) dollars. Also, to avoid topcoding issues and reduce the problem of earnings nominally falling below minimum wage, the top and bottom 5 percent of the values are trimmed out within each cell. Because the trimming is based

Mark S. Handcock is a statistician and Martina Morris is a sociologist at Pennsylvania State University, and Annette Bernhardt is a sociologist at the Institute on Education and the Economy, Teachers College, Columbia University.

on the percentiles within cells rather than across the entire sample, the cells are the unit of analysis. As in the earlier paper, for the regression analyses, the CPS and NLSY79 samples are restricted to respondents who were aged 20 years or older in the survey year and whose earnings and number of weeks worked during the previous calendar year both were positive. The dependent variable is the within-cell standard deviation of trimmed real log annual earnings in the year prior to the interview.

Updating the Gottschalk-Moffitt analysis beyond 1988 requires some changes to the sample selection criteria due to changes in survey coding procedures that have taken place since then. In addition, to focus the sample more tightly on a homogeneous set of white males, some new exclusions are adopted. The following tabulation compares the sample selection criteria used in the present analysis with those used by Gottschalk and Moffitt in their study.

Criteria	*Gottschalk-Moffitt*	*Updated analysis*
Years	1979–88	1979–94
Age range	16–21 in 1979	16–21 in 1979
Race	White	White, non-Hispanic
Enrollment	Employment status recode-based exclusion	No student exclusion
Earnings	Positive	Positive
Regression sample:		
Age	20 years and older	20 years and older
Weeks worked	Positive	Positive

The most important difference in the criteria used here concerns the exclusion of students. On the basis of the "employment status recode" variable, Gottschalk and Moffitt exclude CPS and NLSY79 respondents who reported school attendance as their major activity during the survey week. But the coding for this variable in the CPS was changed in 1988 and it no longer identifies school attendance as a unique status. To preserve consistency across the time series, therefore, this analysis does not directly exclude students in this way. The overall impact of the change is relatively small, though, because several of the other exclusions (positive earnings and number of weeks worked, for example) capture much of the same population.[7]

For each data set, descriptive regression analyses similar to those used in the earlier study were conducted to compare the trends in earnings across the different samples. Let y_{at} be the standard deviation of the log annual wages for workers age a in year t. The model fit by Gottschalk and Moffitt is a simple linear specification:

$$y_{at} = b_0 + b_1 a + b_2 t + e_{at} \qquad a = 20,...,36;\ t = 79,...,94 \qquad [A]$$

where b_1 and b_2 are the coefficients for the linear effects of age and year, respectively. The present analysis extends the earlier study in two ways. First, the regression model is respecified and two alternative specifications are examined: a nonparametric model for the age term and a random-effects model to capture the longitudinal sample dependence in the NLSY79.

The regression residuals for model A show a marked curvilinear pattern in age that is roughly parabolic in nature. The time trend is of primary interest here, rather than the effects of age. Given the correlation between year and age in these samples, however, the age effect must be specified properly to obtain an accurate estimate of the time trend. As the linear age specification compromises the interpretation and statistical significance of the coefficients of both linear coefficients, the model is respecified using a nonparametric age effect, as follows:

$$y_{at} = a_e + b_a + bt + e_{at} \qquad a = 20,...,36;\ t = 79,...,94 \qquad [B]$$

where b_{20}, ..., b_{36} are coefficients for each age and b is the regression parameter for the linear time trend.

It is important to note that the two previous studies have treated both the CPS and the NLSY79 as cross-sectional surveys, although the latter is a longitudinal survey. There are eight cohorts in the NLSY79, defined by respondent's age in 1979, and each cohort is followed across the entire 16 years of the series. Observations from the same cohort in the NLSY79 are likely to be correlated across time, a fact not taken into account in the Gottschalk–Moffitt analysis, the study by Thomas MaCurdy and others (cited earlier), or in the models (A and B) shown above. The cohort sample dependence can be modeled in one of two ways—as a fixed effect or as a random effect. Adding a fixed effect to either model A or model B is not possible because the parameters for age, year, and cohort are perfectly confounded (cohort = year minus age). A random-effect specification is therefore required and also is more appropriate from a substantive standpoint. The interest here is not in the cohort effects as indicators of inherent differences among specific age–year groups. The cohorts are simply samples from their populations, and this study seeks to capture the covariance in these samples over time, rather than an estimate of a cohort-specific level effect. Therefore, model B is respecified for the NLSY79 to include a random effect for cohort, as follows:

$$y_{atc} = a + b_a + bt + e_{tc}$$
$$a = 20,...,36;\ t = 79,...,94;\ c = 1,...,8; \qquad [C]$$

$$e_{tc} = f_c + s_{tc}$$

where b_{20}, ..., b_{36} are coefficients for each age, b is the coef-

ficient for the linear effect of year, and f_1, ..., f_8 are random variance components for each cohort. Because it requires no assumptions about the parametric form of the random cohort effects, a generalized estimating equation (GEE) is used to fit the model.[8]

For all of the linear models, weights are used to reflect the differing variances of the y_{at} component of the model.[9] In the GEE models, the variance-covariance weight matrix includes covariance estimates in the off-diagonal cells to adjust for the longitudinal cohort sample dependence. All models are fit using the ***S-PLUS*** statistical program.[10]

The second way in which the present study extends the Gottschalk–Moffitt analysis is by reexamining the discrepancies in earnings dispersion by labor force status. Gottschalk and Moffitt use several indicators as proxies of labor force attachment in an attempt to explain the discrepancy in earnings trends: the employment status recode variable, more than 40 weeks worked in the past year, and age 23 years and older (presumably to exclude most college-age students). The present study takes a more direct approach, subdividing the sample into two groups: full-time, year-round workers (FTFY) and others (non-FTFY). The FTFY group comprises those who worked 35 or more hours per week and 50 or more weeks per year during the previous calendar year; the non-FTFY group comprises those who had positive earnings and hours worked but who did not work full time and year round. For the CPS, the constructed variable that identifies this status is used, and for the NLSY79, hours and weeks are selected directly. The definition is the same in both samples. The idea here, as in the earlier study, is to compare workers with relatively strong attachments to the labor force with workers who are less attached to the labor force.

Results

Tables 1 and 2 provide summary statistics for labor force attachment and annual earnings for workers in both data sets in 1979, the first year of the series. The sample selections reflect the updated analysis criteria and can be compared with the corresponding tables in the paper by Gottschalk and Moffitt. Table 2 shows patterns similar to those found in the earlier study—a significantly larger portion of the NLSY79 sample reports working 40 weeks or more per year. While fairly pronounced in 1979, this discrepancy in the number of weeks worked during the year declines in subsequent years.

Despite the difference in reported number of weeks worked, the earnings figures in table 2 are quite similar across the two samples. There are no systematic differences in either means or variances. The numerical values are different than those reported by Gottschalk and Moffitt, due largely to the inclusion here of students who had been excluded in the earlier study on the basis of the employment status recode variable.[11] The bottom portion of the table shows the statistics for FTFY respondents—a group likely to exclude such students—and here the two samples become very close.

The trends in earnings variances over time for the two samples are shown in chart 1. They show a general decrease in earnings dispersion with age, and this pattern is much stronger than the trend over time within specific age groups. The NLSY79 estimates are more variable, reflecting the smaller sample sizes. Net of the differences in variability between the two samples, the greatest differences between them occur within the younger age groups—those aged 20 to 24 years. These differences are not very systematic, and in particular, they do not appear to take the form of consistently stronger increasing trends over time in the CPS. There is some conver-

Table 1. Basic descriptive statistics for 1979 survey year

Age	High school graduates (in percent)		Unweighted *N*		Among high school graduates				Percent working full time, year round	
					Percent working at least 1 week during the year		Percent working 40 or more weeks during the year			
	NLSY79	CPS	NLSY79	CPS	NLSY79	CPS	NLSY79	CPS	NLSY79	CPS
Total (all ages)	44.7	57.8	796	3,261	95.6	92.5	52.9	48.8	26.1	28.4
16	0.4	0.2	1	4	...	...	...	...	...	...
17	0.9	0.5	4	30	...	75.9	...	19.0	...	3.4
18	45.6	47.5	145	507	96.3	90.3	39.2	37.3	10.5	15.3
19	79.3	80.3	218	903	95.2	91.7	49.0	43.4	23.5	22.6
20	86.7	88.4	224	885	94.4	93.5	62.3	50.1	36.4	32.4
21	87.4	87.5	204	932	96.6	93.9	54.6	59.1	27.2	37.1

Table 2. Basic income statistics for survey year 1979

Age	Unweighted *N*		Income mean		Coefficient of variation		Log income mean		Standard deviation	
	NLSY79	CPS	NLSY79	CPS	NLSY79	CPS	NLSY79	CPS	NLSY79	CPS
All workers										
16	1	2	…	1,221	…	…	…	7.11	…	0.00
17	2	28	2,608	3,071	463	3,323	7.84	7.54	.35	1.08
18	118	601	3,814	4,163	1,416	1,621	8.02	8.12	.73	.69
19	198	1,100	6,120	5,819	2,716	2,817	8.45	8.39	.80	.80
20	214	1,160	8,373	6,643	2,661	2,938	8.84	8.53	.67	.80
21	202	1,230	8,812	8,991	3,793	3,768	8.82	8.83	.79	.81
Full-time, year-round workers										
16	0	0	…	…	…	…	…	…	…	…
17	0	1	…	3,497	…	…	…	8.16	…	…
18	12	93	5,380	7,547	839	805	8.48	8.87	.55	.36
19	45	245	10,067	10,414	1,354	1,012	9.13	9.18	.47	.42
20	83	385	11,413	10,823	1,607	1,066	9.24	9.23	.51	.38
21	51	481	13,648	13,374	1,466	1,581	9.46	9.42	.39	.48

NOTE: Statistics are calculated using sample weights and 5-percent trim of top and bottom earnings. Unweighted *N* reflects post-trim cell values.

gence between the two samples for the older respondents, but the earnings dispersion for the NLSY79 is about 10 percent lower, on average, than for the CPS. By contrast, the cell median incomes in the NLSY79 are consistently about 20 percent higher than the corresponding CPS cell means (data not shown here). Once the two samples of respondents settle into their prime working years, then, the annual earnings reported in the NLSY79 are both higher and less variable than those reported in the CPS.

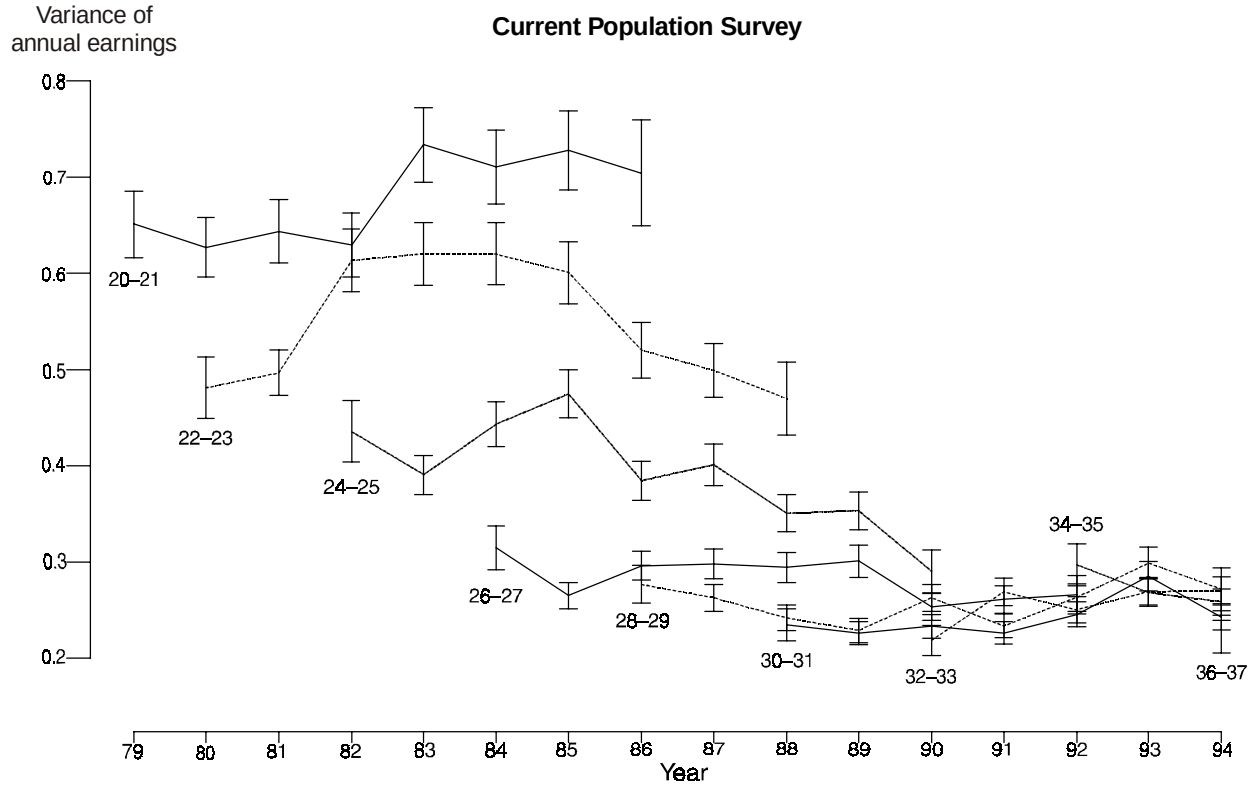
The standard deviations are modeled by reverting to cells defined by survey year and single year of age. Much like the Gottschalk–Moffitt study, attention here is restricted to those aged 20 years and older, with positive weeks worked in the previous calendar year. The results are displayed in table 3. All coefficients are multiplied by 10 to be consistent with the values reported by Gottschalk and Moffitt. The coefficients can be interpreted as the change in standard deviation over a 10-year period.

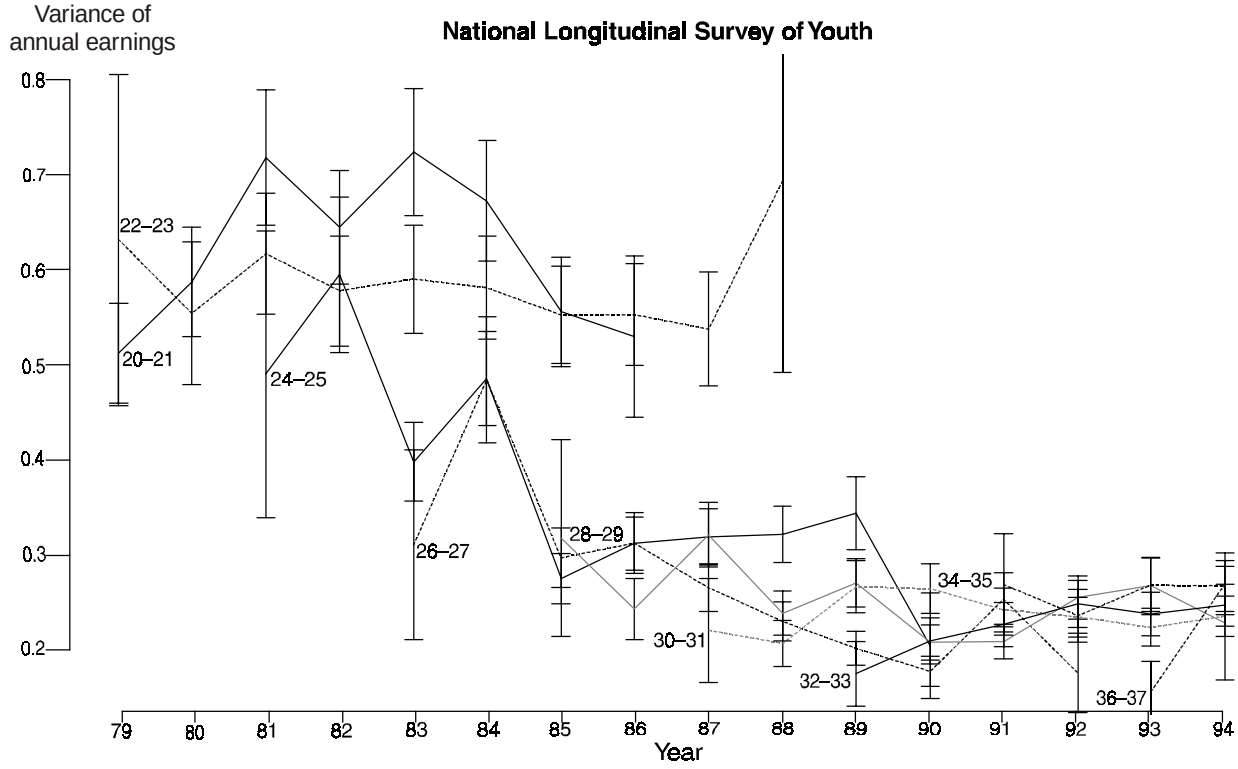
The results obtained by Gottschalk and Moffitt are shown in the first three rows of the table for comparison. Consider first their results based on the employment status recode schooling exclusion. For the CPS, they find a positive but not significant upward trend in earnings dispersion, while the corresponding trend for the NLSY79 is negative and also not significant. Using a more specific measure of school enrollment over the past year that is available in the NLSY79 to exclude students in that sample, they find the coefficient for the trend in dispersion changes sign and becomes as strongly positive as it had been negative, though still not significant. Further restricting this NLSY79 sample to those aged 23 years and older, they find the coefficient changes sign again and is now much more strongly negative than it had been, though still not significant.

The Gottschalk–Moffitt estimate of the time trend is thus extremely sensitive to the sample exclusions. The same is true in the present analysis, in part due to the relatively small number of observations in each cell after the screens for positive earnings and weeks worked and the 10-percent trimming. This makes for a high level of instability in the cell-specific estimates of the earnings variance, and these in turn have a large impact on the within-age trend estimates. The latter is due to the interaction between the model, which estimates the time trend within age, and the structure of the sample. While the two surveys cover 16 years, age groups are observed for, at most, 8 years, and the average for persons aged 20 years and older is 6.3 years. The moving cohort window is thus not an ideal structure for capturing trends within age over time. When drawing inferences about the discrepancies between the two samples, it should be kept in mind that the estimates are not particularly robust.

The remaining rows in table 3 present the results from the updated analysis. In the first set, we restrict the sample to the years used by Gottschalk and Moffitt, 1979–88. The differences between the results for model A and the results in the first row of the Gottschalk–Moffitt figures reflect the difference in the sample restrictions between the two analyses—namely, the inclusion in this analysis of students who were excluded from the earlier study on the basis of the CPS employment status recode, as well as the exclusion here of Hispanics. The impacts are not dramatic, with the CPS coefficient becoming slightly less positive under the new sample restrictions. The NLSY79 coefficient

Chart 1. **Variance of log annual earnings for employed high school graduates by age/year cells, Current Population Survey and National Longitudinal Survey of Youth, 1979 94**





NOTE: Respondents are grouped in 2-year intervals. Bars show the 95-percent confidence intervals.

becomes more negative and now also is statistically significant, though in magnitude it still lies within the range of estimates reported in the earlier study.

When a nonparametric specification for age is adopted in model B, the discrepancy declines—the CPS coefficient increases modestly, and the NLSY79 coefficient becomes much less negative. When the random effect for the longitudinal cohort dependence in the NLSY79 (model C) is added, the coefficient for the time trend again becomes slightly less negative, and now it is about 30 percent lower than the initial estimate in model A. While the numerical results obtained in the earlier study are not replicated exactly, the general pattern is replicated, showing an increasing trend for earnings dispersion in the CPS and a decreasing trend for the NLSY79 . The magnitude of the discrepancy and of the negative trend in the NLSY79 becomes smaller in both of the respecified models.

The next set of results shown in table 3 (labeled all workers) updates the analysis to 1994. For the CPS, the trend in earnings dispersion is now significantly negative in model A, as is the trend for the NLSY79. With the nonparametric age effect, the sign of the CPS coefficient changes to become positive (although weakly so and not significant), while the magnitude of the NLSY79 coefficient is still negative but reduced by about half. Adding the random effect to the NLSY79 slightly increases the magnitude of the negative trend, but it is still 40 percent lower than the estimate under the initial model. Respecifying the model once again reduced the discrepancy between the two samples.

The results from model C are graphically displayed in chart 2. The top panel plots the nonparametric age-effect estimates. The results show that earnings dispersion is highest among the young, and it falls steeply through the mid-twenties age groups. For the NLSY79, dispersion then begins to rise slightly, while for the CPS, the decline continues through the early-thirties age groups, though less steeply, and then also begins to rise. The nonlinearity for the NLSY79 is more pronounced, which helps to explain why the nonparametric specification in Model B has a relatively larger impact on the trend coefficient for that sample.

The bottom panel of chart 2 shows the partial regression plot of earnings dispersion by year after adjusting for age. The trend lines are nonparametric local-linear estimates. As can be seen, the CPS trend is modestly positive. The plot for the NLSY79, by contrast, clearly shows a negative trend. Note, however, the large residual variation. The magnitudes of the time trends for both samples are modest relative to the residual variability.

Next, the analysis is restricted to full-time, year-round workers in order to determine whether the discrepancies in earnings dispersion between the two samples persist among the core group of workers with the strongest attachment to the labor force. This group becomes an increasingly larger share of the two samples over time, rising from about 35 percent of the regression-eligible sample in 1979 to 80 percent in 1994. If the trend differential persists for these workers, then it is a fundamental and pervasive discrepancy. If not, then the samples are comparable for the core workers, and some progress has been made in narrowing down the possible sources of the problem.

Table 3. Regression results

Sample restriction and model	CPS	NLSY79
Gottschalk-Moffitt analysis:		
CPS—not in school	0.019	–0.038
NLSY79—nonenrolled	…	.038
NLSY79—23 years and older	…	–.100
Updated analysis:		
1979–88 only		
A	.015	[2]–.124
B	.020	[2]–.093
C	…	[2]–.089
All workers, 1979–94		
A	[1]–.049	[2]–.165
B	.009	[2]–.085
C	…	[2]–.092
Full-time year-round workers, 1979–94		
A	[2].025	[1]–.030
B	[2].032	–.020
C	…	[2].036
Part-time, part-year workers, 1979–94		
A	.030	[1]–.126
B	[1].042	–.096
C	…	[2]–.116
Full-time, year-round workers, 1979–94, excluding self-employed		
A	[2].033	–.019
B	[2].041	–.004
C	…	.027

[1] Statistically significant at the 10-percent confidence level.
[2] Statistically significant at the 5-percent confidence level.

NOTE: Model A specifies linear effects for both age and year, model B specifies a nonparametric age effect, and Model C includes a random effect for longitudinal cohort dependence in the NLSY79.

The trend coefficient under model A reproduces the discrepancy observed above, but the negative trend for the NLSY79 is substantially smaller than in all of the previous analyses. The estimates from model B are consistent with the earlier pattern—that is, the discrepancy narrows as the trend becomes more positive for the CPS and less negative for the NLSY79. When the random effect for the sample dependence in model C is added, however, the NLSY79 coefficient changes sign, becoming strongly positive and similar in magnitude to the CPS coefficient, though not statistically significant. Under model C, then, both samples of full-time, year-round workers show a

positive trend in earnings dispersion of comparable magnitude.

The results for the other (non-FTFY) workers show the opposite pattern, with the discrepancy very large under model A and virtually unchanged under model C. For these workers, opposite trends are seen in earnings dispersion for the two samples—dispersion grows over time in the CPS, while it declines over time in the NLSY79. The pattern of statistical significance is also different for this subgroup, with the NLSY79 trends testing highly significant and the CPS trends testing only modestly significant.

The age effects and partial regression plots for model C for the full-time, year-round workers and for the other workers are shown in chart 3. The pattern of higher dispersion for older NLSY79 respondents also is visible here in both subgroups. The smoothed trend lines are clearly different, however, with the FTFY workers in both the CPS and NLSY79 samples now showing a weak positive trend. The residual variability also differs: it is now lower for the FTFY workers and higher for the non-FTFY workers. The smoothed trend lines do not tell an entirely unambiguous story—when the endpoints are excluded, a different trend sometimes emerges. The regression line would be even more strongly influenced by the high leverage points at the extremes, simply reinforcing the earlier point that caution is appropriate when drawing inferences from any of the trend coefficients estimated from these samples.

One final analysis was conducted in which the self-employed were excluded. This is a group known to have highly variable earnings. They are almost universally excluded in studies of earnings inequality because their earnings determination process is fundamentally different from that of wage and salary workers. Excluding the self-employed, the pattern obtained is basically the same as that of the full sample of FTFY workers: in the final specification of model C, both samples again show a positive trend of similar magnitude in earnings dispersion over time.

These analyses suggest that the earnings dispersion discrepancy found by Gottschalk and Moffitt results largely from the specification of their regression model as well as a trend that appears to be driven by workers who do not work full time and year round. To examine the latter, chart 4 shows the trends in earnings dispersion by age-year cell separately for FTFY and non-FTFY workers.[12] The trends for FTFY workers look similar for the two samples—that is, both groups show a modest upward trend. The age effects discussed earlier (see chart 1) are completely absent here. In the graph for non-FTFY workers, by contrast, the CPS shows a fairly stable pattern of earnings dispersion over time, while the trend for the NLSY79 is somewhat negative. This clearly is what is driving the negative trend in the NLSY79 data when both groups of workers are combined. For non-FTFY workers, the age differences are absent as well. Thus, what at first appears to be an age effect in the graph for all workers actually is a composition effect—as age increases, the majority of workers shift from non-FTFY status to working full time and year round.

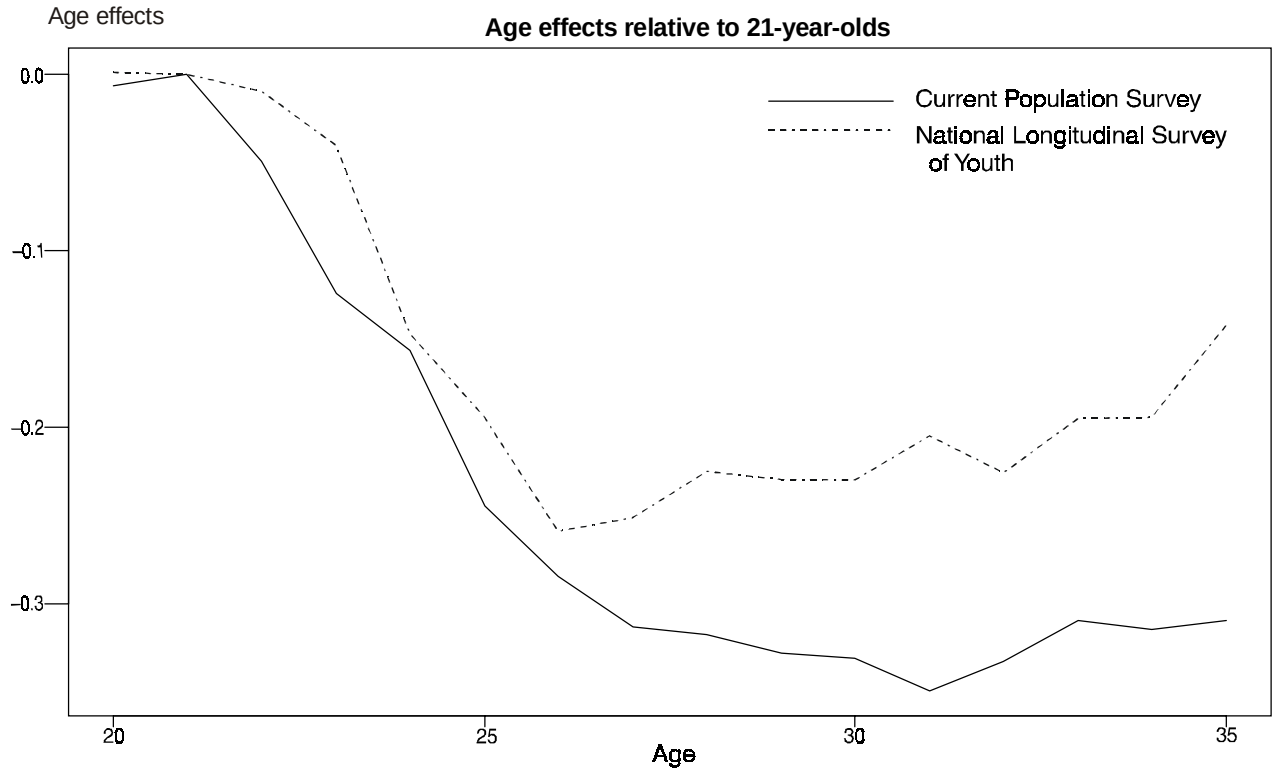
To better understand the nature of these discrepancies, it is useful to look at estimates of the distributions themselves. Chart 5 shows the 1979 earnings densities for the two samples as an example.[13] The top panel corresponds to all workers. While the two distributions are similar at the higher earnings levels, the CPS sample has a longer, denser lower tail than the NLSY79 sample. The bottom panel shows the corresponding distributions for non-FTFY workers. The CPS distribution is strongly downshifted, indicating lower levels of reported earnings compared with the NLSY79, and the bottom tail of the distribution for these workers reaches much further down the earnings scale. The location of the lower tail of the non-FTFY earnings density, from about 6 to 8 on the log scale, corresponds exactly to the location of the lower tail differences in the distribution for all workers. The plot for FTFY workers, not shown here, looks much like the plot for all workers, without the greater relative density in the lower tail of the CPS.

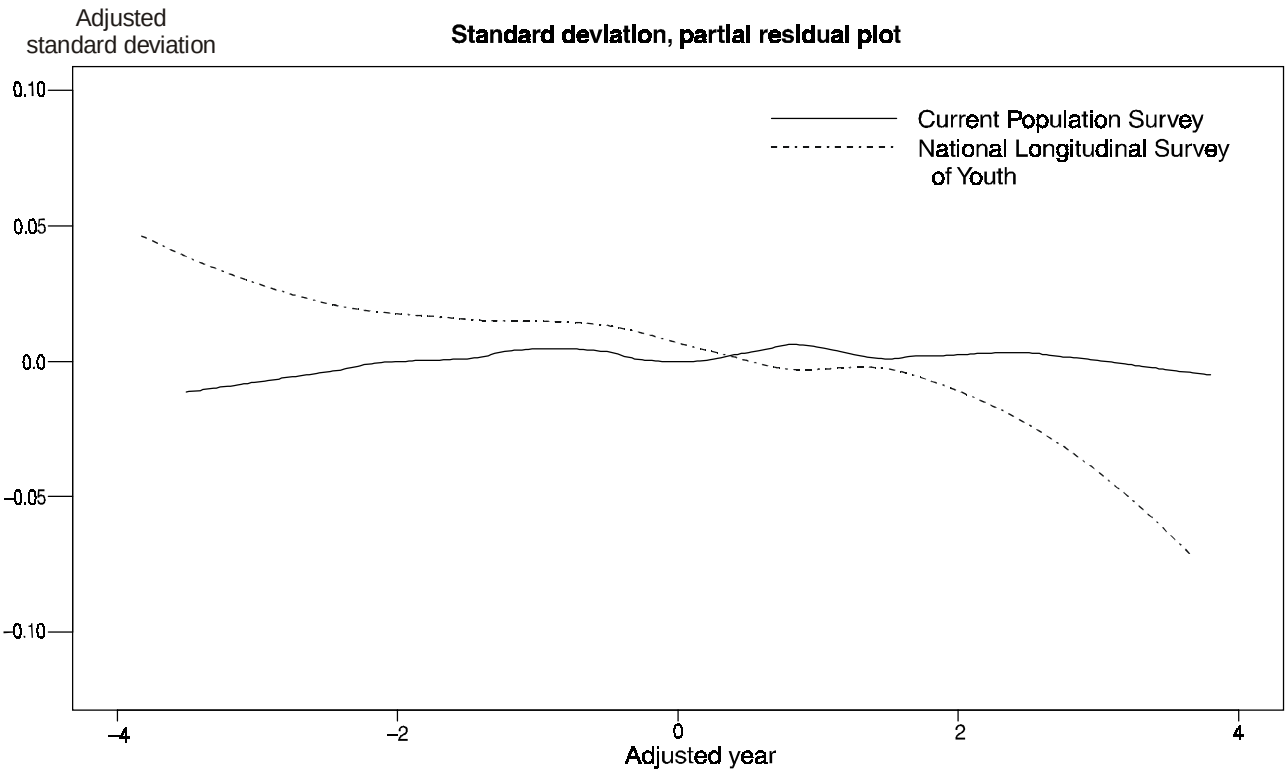
This lower tail discrepancy becomes more pronounced over time, as can be seen by the 90:50 and 50:10 earnings ratios for non-FTFY workers shown in chart 6. The 50:10 ratio for the two samples is relatively similar at the start of the series, but the CPS ratio increases over time while the NLSY79 ratio declines. Given the consistently lower median reported earnings in the CPS, the rise in the 50:10 ratio implies an increasingly longer tail at the bottom of the distribution than that observed in the NLSY79. The 90:50 ratios are more similar for the two samples, with both showing a downward trend over time, though the timing of the decline is different. The variance differential between the two samples is thus being driven primarily by the discrepancies in the lower tails. Specifically, it is being driven by the longer lower tail of the CPS non-FTFY earnings distribution.

Discussion

The discrepant findings in the trends in annual earnings dispersion between the CPS and the NLSY79 appear to be a function of the model specification and the non-FTFY workers. Regression diagnostics clearly show that a linear specification for age is not appropriate, and fitting a nonparametric effect reduces the discrepancy in the estimated dispersion trends by one-third to one-half. Treating the two samples as cross-sectional, thus ignoring the longitudinal cohort dependence in the NLSY79, also is not appropriate. Modeling the cohort dependence in the NLSY79 changes the estimates of the dispersion trend, especially when the sample is restricted to FTFY workers.

Chart 2. **Estimated age effects, regression results under model C**

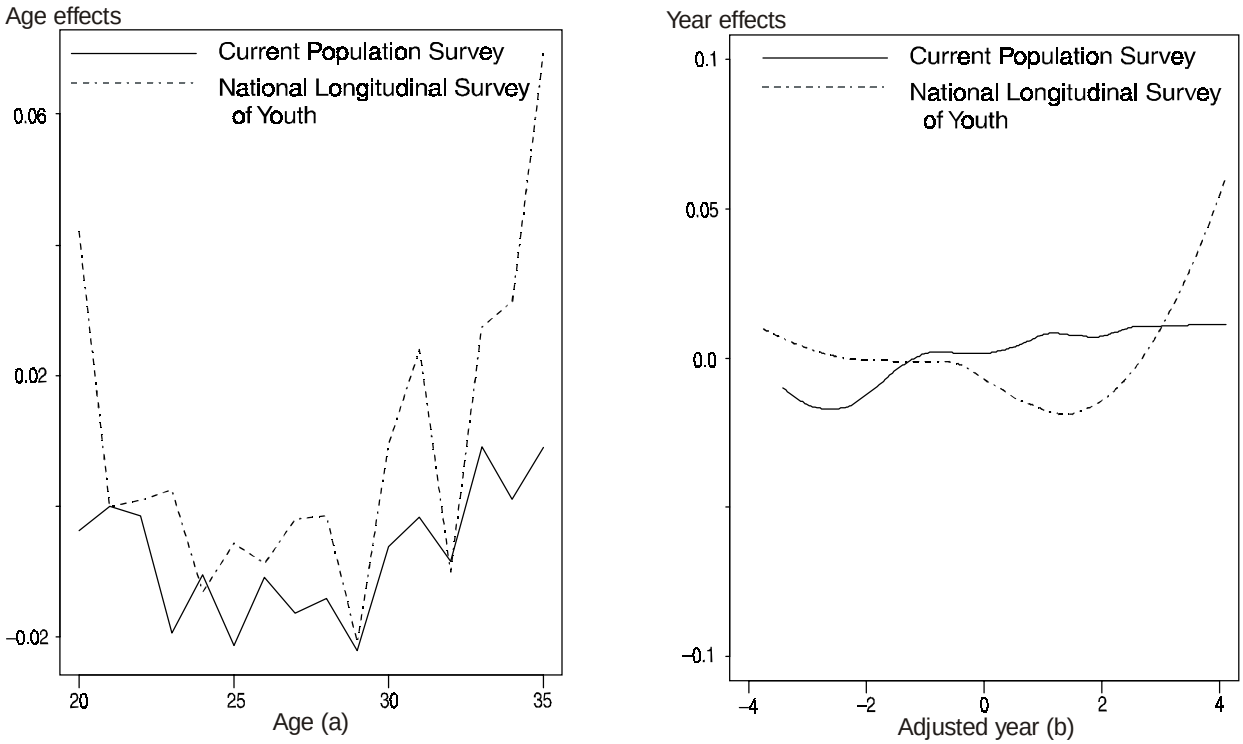




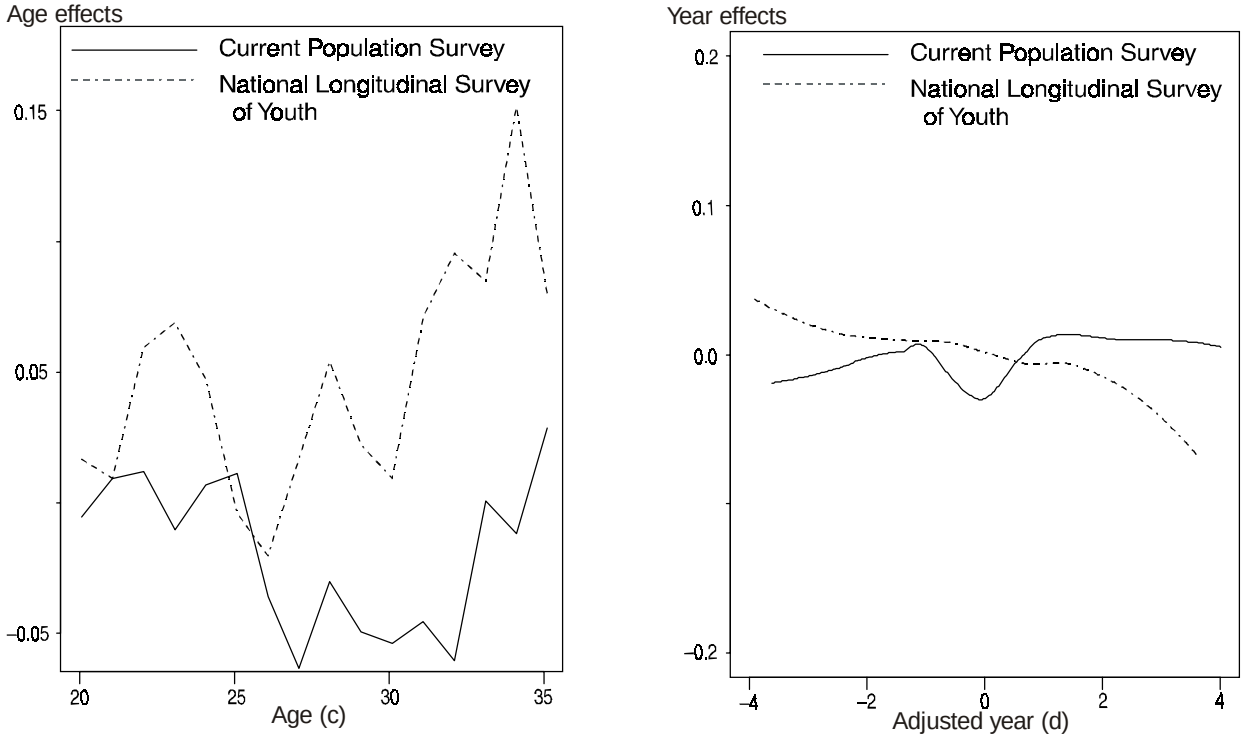
NOTE: Regression results are for both full-time, full-year and non-full-time, -full-year workers. Trend shown as a local linear smoothed estimate.

Chart 3. **Estimated age effects relative to 21-year-olds by full-time/part-time status, and standard deviation partial residual plot, regression results under model C**

Full-time, full-year workers



Part-time or part-year workers



NOTE: Trend shown in panel (d) as a local linear smoothed estimate.

Chart 4. Variance of log annual earnings by age/year cells, 1979 94

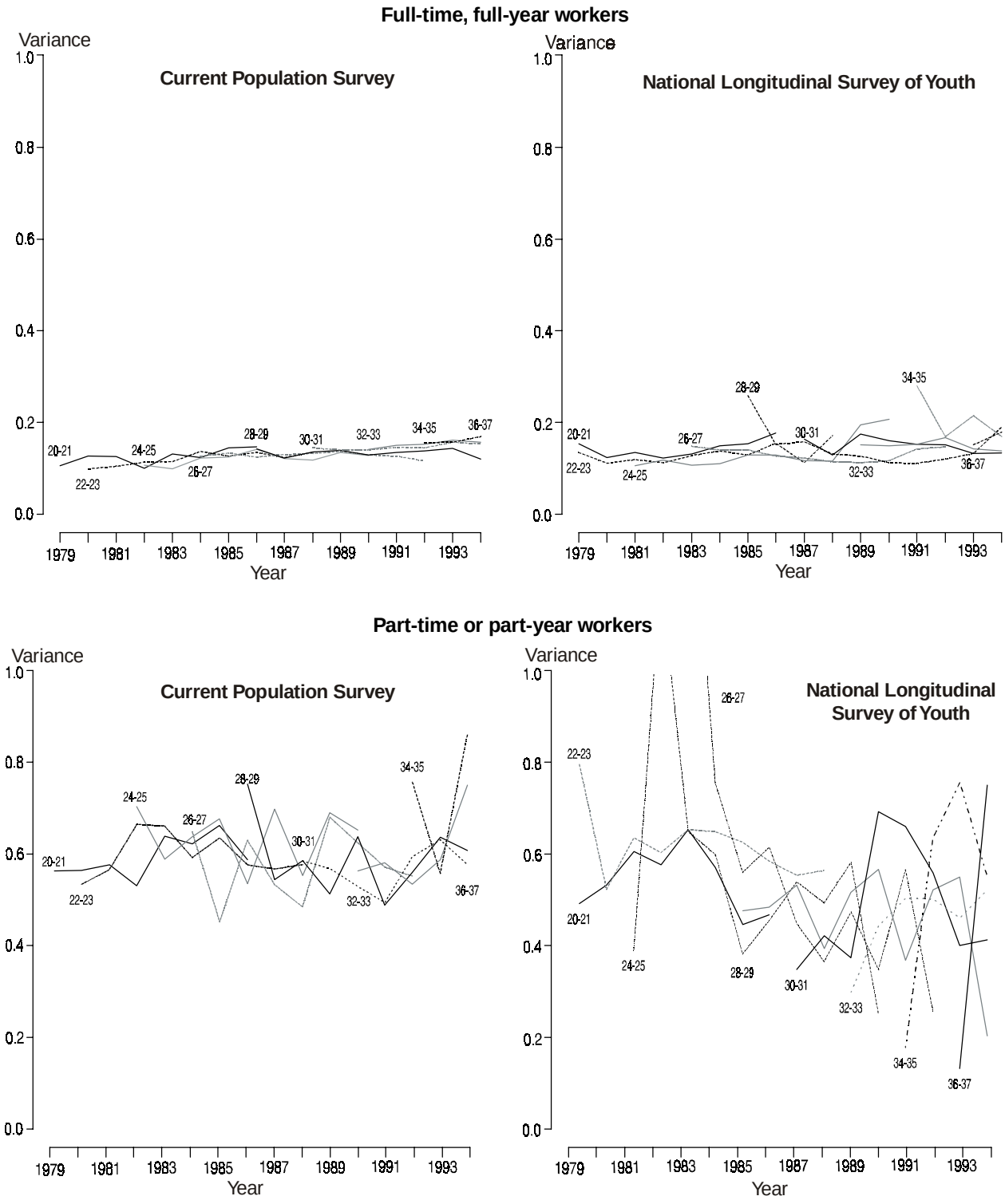


Chart 5. **Annual earnings density estimates, 1979**

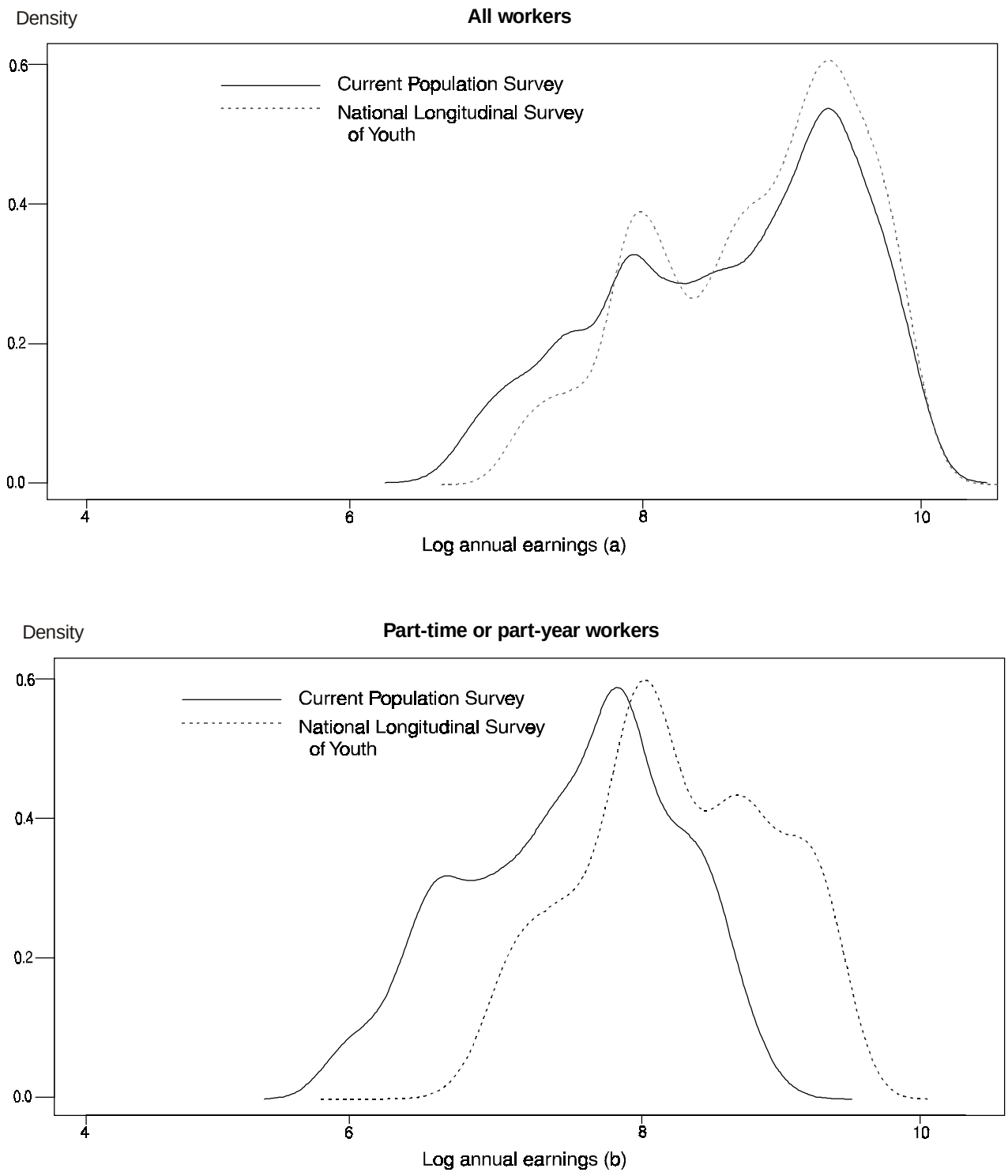
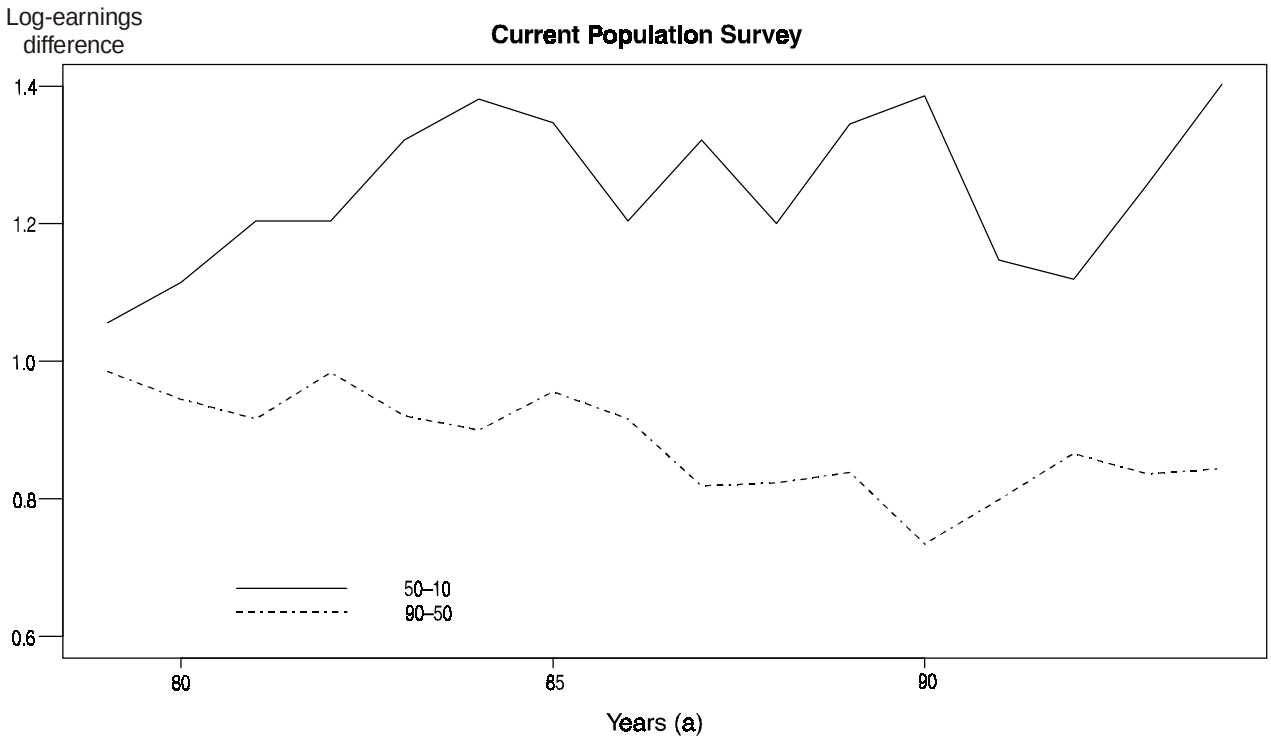
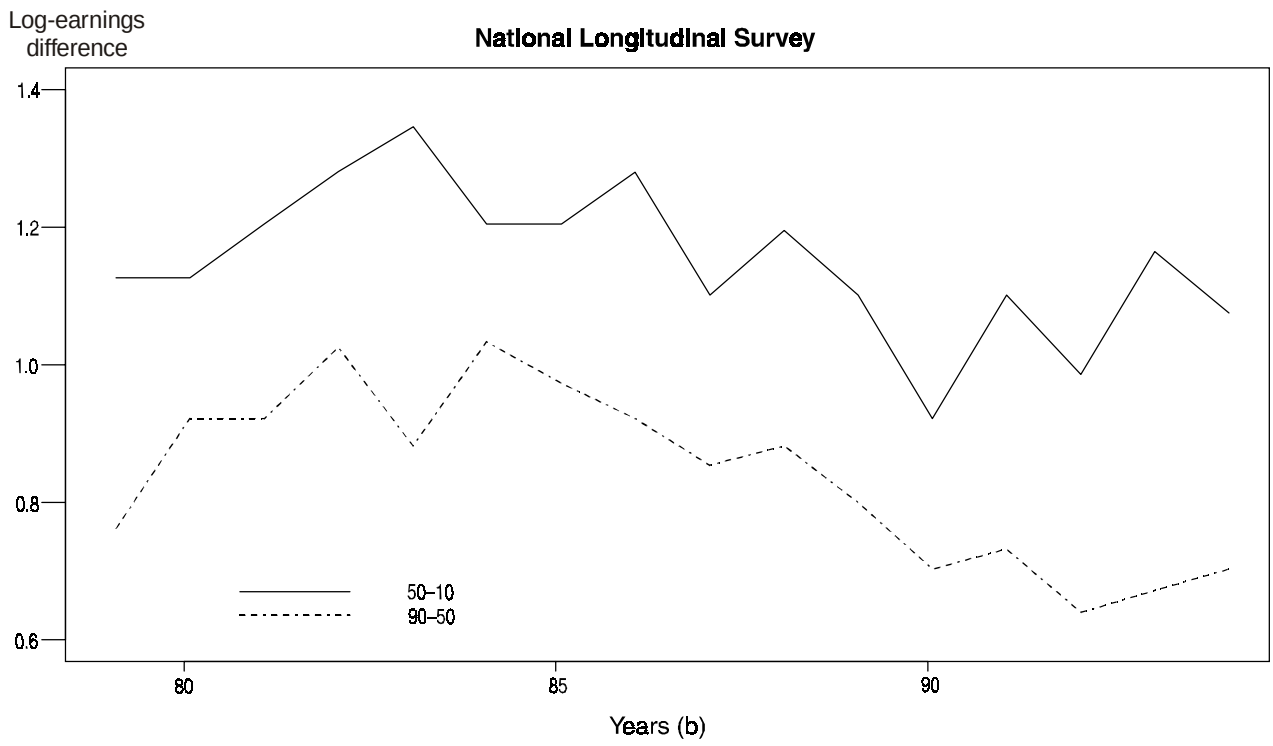


Chart 6. **Trends in the 90:50 and 50:10 annual earnings ratios for part-time or part-year workers, 1979 94**





After these corrections, the earnings dispersion trends for FTFY workers look remarkably similar for the two samples. Formal analysis confirms this visual impression—the estimated trends in earnings dispersion are nearly identical. Thus, restricting the samples to FTFY workers, no significant discrepancy in earnings variance is found between the two data sets: both the CPS and NLSY79 show a general trend of increasing earnings dispersion over time.

The trends in earnings dispersion among non-FTFY workers, however, appear to be different in the two samples. Closer examination of the two earnings distributions shows clearly that the distribution of reported annual earnings among non-FTFY workers in the CPS is both strongly downshifted and skewed more to the left than in the NLSY79. CPS respondents who do not work full time and year round not only report lower earnings, on average, but also the bottom tail of their distribution reaches much farther down the earnings scale. These differences already are pronounced in 1979, and they grow over time, thus contributing directly to the growing discrepancy between the two samples.

For both groups of workers, annual earnings reports are higher in the NLSY79 than in the CPS by about 20 percent at the median. This begins to suggest that the primary source of the discrepancy may be underreporting in the CPS. The most likely explanation is differences in the respective questionnaires, because neither sample bias nor attrition bias has been suggested as a problem in the NLSY79.[14] As noted in the study by Gottschalk and Moffitt, the design of the NLSY79 questionnaire probably increases the accuracy of earnings reports. The sequence of questions asked about individual jobs in the NLSY79 aids in the recall of both earnings and hours relative to the CPS, and the effect would be expected to be strongest for part-time or part-year workers with irregular schedules and sources of earnings.

In addition, the NLSY79 is administered as a face-to-face interview, whereas the CPS, except for the initial interview, usually is administered by telephone.[15] This probably will raise the validity and reliability of the NLSY79 data relative to the CPS. The longitudinal basis of the NLSY79 provides a continuing relationship between the respondents and the survey organization. The promise of confidentiality has been met over time, and respondents may feel more comfortable disclosing sensitive information on earnings. Also, in the CPS, proxy reports may be a factor. All of this suggests that the discrepancies in non-FTFY annual earnings reports between the CPS and the NLSY79 may be due to underreporting in the CPS.

It is worth reiterating, however, that the regression trend estimates obtained from these samples should be interpreted with care. They were found to be highly sensitive to small changes in sample selection and model specification. The structure of the analytic question, which focuses analysis on the trends within age over time, leads to both relatively small cell sizes for estimating dispersion, and a mismatch between sample structure and the analytic task. To obtain stable estimates of the time trend, one would need relatively long periods of observation within age groups. The cohort scheme of the NLSY79, with its 8-year moving age window over time, only provides a maximum of 8 years during which any respondents are observed at a particular age, and some of the age segments include less than 2 years of observation.[16] Of course, the equivalent CPS sample reflects the same constraints. While the goal of benchmarking the NLSY79 against the CPS is an important one, the NLSY79 sample structure is not ideal for answering the question posed here, and it is not clear that the survey would ever be used in this fashion.

With that caveat, however, the findings described in this article still attest to the validity of the NLSY79 data. Researchers should therefore take advantage of these data to examine the longitudinal questions for which this survey was designed. In general, the National Longitudinal Surveys, with their unique employer identification codes, remain the only longitudinal data set with an accurate measure of job and employer stability—a significant feature, given the many contradictory empirical findings in this field.[17] The age range covered by the survey provides a detailed window into the period when roughly two-thirds of lifetime job changes and wage growth occur.[18] These also are the formative years of labor market experience when long-term relationships with employers are established. The two National Longitudinal Survey cohorts also bracket the growth in earnings inequality that emerged in the 1980s. Together, the cohorts of the National Longitudinal Surveys provide a unique resource for the analysis of these and other important economic and social issues covering the last 30 years. □

Notes

[1] The Panel Study of Income Dynamics (PSID), begun in 1968, is conducted by the Survey Research Center, Institute for Social Research, University of Michigan. The PSID is a longitudinal study of a representative sample of U.S. individuals (men, women, and children) and the family units in which they reside. It emphasizes the dynamic aspects of economic and demographic behavior, but its content is broad, including sociological and psychological measures. As a consequence of low attrition rates and the success of recontact efforts, the sample size has grown dramatically in recent years, from about 7,000 core households in 1990 to almost 8,700 in 1995. As of 1995, the PSID had collected information about more than 50,000 individuals spanning as much as 28 years of their lives. For more information on the PSID, visit their website at **http://www.isr.umich.edu/src/psid/**.

[2] The National Longitudinal Surveys (NLS), sponsored and directed by the Bureau of Labor Statistics, gather detailed information about the labor market experiences and other aspects of the lives of six groups of men and women. Over the years, a variety of other government agencies, such as the National Institute of Child Health and Human Development, the Department of Defense, and the Department of Education, the Department of Jus-

tice, the National Institute on Drug Abuse, and the National School to Work Office, have funded components of the surveys that provided data relevant to their missions. As a result, the surveys include data about a wide range of events such as schooling and career transitions, marriage and fertility, training investments, child-care usage, and drug and alcohol use. The depth and breadth of each survey allow for analysis of an expansive variety of topics such as the transition from school to work, job mobility, youth unemployment, educational attainment and the returns to education, welfare recipiency, the impact of training, and retirement decisions.

The first set of surveys, initiated in 1966, consisted of four cohorts. These four groups are referred to as the "older men," "mature women," "young men," and "young women" cohorts of the NLS, and are known collectively as the "original cohorts." In 1979, a longitudinal study of a cohort of young men and women aged 14 to 22 was begun. This sample of youth was called the National Longitudinal Survey of Youth 1979 (NLSY79). In 1986, the NLSY79 was expanded to include surveys of the children born to women in that cohort, with the new cohort called the NLSY79 Children. In 1997, the NLS program was again expanded with a new cohort of young people aged 12 to 16 as of December 31, 1996. This new cohort is the National Longitudinal Survey of Youth 1997 (NLSY97).

The National Longitudinal Surveys, especially the NLSY79, have exceptional retention rates. As a result, many NLS survey members have been followed for many years, some for decades, allowing researchers to study large panels of men, women, and children over significant segments of their lives. For more information on the National Longitudinal Surveys, see the *NLS Handbook, 1999* (Bureau of Labor Statistics, 1999).

[3] See Peter Gottschalk and Robert A. Moffitt, "Earnings and wage distributions in the NLS, CPS, and PSID," final report to the U.S. Department of Labor (Brown University, 1992).

[4] The Current Population Survey (CPS), which uses a scientifically selected sample of about 50,000 households, is conducted monthly for the Bureau of Labor Statistics by the Bureau of the Census. The CPS provides statistics on the labor force status of the civilian noninstitutional population of the United States, aged 16 years or older. In the CPS, respondents are asked about their activity during the week that includes the 12th day of the month, the so-called reference week. As such, the CPS is a cross-sectional survey of the population, as opposed to a longitudinal survey like the NLS. For more information on the CPS, see *BLS Handbook of Methods*, Bulletin 2490 (Bureau of Labor Statistics, April 1997), pp. 4–14.

[5] See Thomas MaCurdy, Thomas Mroz, and R. Mark Gritz, "An Evaluation of the National Longitudinal Survey of Youth," *Journal of Human Resources*, spring 1998, pp. 345–436.

[6] To further minimize heterogeneity, this study excludes Hispanics from the samples analyzed. The study by Gottschalk and Moffitt made no such exclusion.

[7] For the regression-eligible sample used here, ESR-type students represent about 15 percent of the respondents in 1979, dropping to 5 percent in 1985 and down to 1 percent by 1988.

[8] See Peter J. Diggle, Kung-Yee Liang, and Scott L. Zeger, *Analysis of Longitudinal Data* (New York, Oxford University Press), 1994.

[9] See Gottschalk and Moffitt, "Earnings and wage distribution in the NLS, CPS, and PSID," p. 7.

[10] *S-PLUS* is an enhanced version of the *S* environment for data analysis. Unix and Windows versions are available from MathSoft, Inc. The programs used for the analysis in this paper are available from the authors.

[11] Nearly a third of the sample in table 2 has an employment status code indicating enrollment in the survey week in 1979, and these respondents would have been excluded in the Gottschalk–Moffitt analysis. As noted earlier, the portion of students with the employment status recode variable is lower for the regression-eligible sample, which is further restricted to those aged 20 years and older who also had a positive number of weeks worked.

[12] As in chart 1, 2-year age groups are used. For FTFY workers, the values average about 180 respondents per cell for the NLSY79 and about 870 respondents per cell for the CPS. For non-FTFY workers, the corresponding values average about 90 and 300, respectively.

[13] For this figure ages within a year are pooled, but the distributions have been compositionally adjusted for the differences in marginal age distributions between the CPS and NLSY79.

[14] See MaCurdy and others, "An Evaluation of the National Longitudinal Survey of Youth."

[15] In the CPS, respondents are part of the survey for 4 consecutive months, then they are out of the survey for the following 8 months, and finally they are back in the survey for 4 more months the following year. The first interviews are supposed to take place in person, at the home of the respondents, although face-to-face interviews are not always possible. In any case, subsequent interviews are conducted by telephone.

[16] Ages 20 to 29 provide 8 years of observation each, other ages in the 16-to-36 year range provide 8 minus the difference to the closer of the two endpoints. In the analysis by Gottschalk and Moffitt, which only included up to survey year 1988, only three ages (20 to 23) would have provided 8 years of observation; all others would have provided fewer years of observation.

[17] See A.D. Bernhardt, M. Handcock, and M. Scott, "Trends in Job Instability and Wages for Young Adult Men," *Journal of Labor Economics, Part 2*, October, 1999, pp. S65–90.

[18] See Kevin Murphy and Finnis Welch, "Empirical Age–Earnings Profiles," *Journal of Labor Economics*, April 1990, pp. 202–29; and Robert Topel and Michael Ward, "Job Mobility and the Careers of Young Men," *Quarterly Journal of Economics*, May 1992, pp. 439–79.