LONG-RUN CONVERGENCE OF ETHNIC SKILL DIFFERENTIALS, REVISITED*

GEORGE J. BORJAS

In my original study, "Long-Run Convergence of Ethnic Skill Differentials," I concluded that the ethnic differentials introduced by the Great Migration seemed to persist into the second and third generations. Alba, Lutz, and Vesselinov argue that my study contained a number of conceptual and data problems, and conclude that the correlation between the skills of the first and the third generations disappears when these problems are taken into account. My reanalysis of the Alba et al. data, however, documents a stronger link between the skills of the first and the third generations than suggested by those authors.

In this issue of *Demography*, Alba, Lutz, and Vesselinov (henceforth ALV) critique the empirical analysis in my 1994 paper "Long-Run Convergence of Ethnic Skill Differentials: The Children and Grandchildren of the Great Migration." In that paper I analyzed microdata from the 1910, 1940, and the 1980 censuses and from the General Social Surveys (GSS) to determine whether the skill differences that existed among the immigrant groups that arrived in the Great Migration were still apparent among their children and grandchildren. I concluded that the ethnic differences seemed to persist into the second and third generations.

ALV question the link in skills between the first and the third generations. They raise a number of conceptual and data problems, and conclude that the correlation between the skills of the first and the third generations disappears when these problems are taken into account. Because the new evidence presented by ALV comes exclusively from the sample of third-generation persons in the GSS, I limit my comments to these data.

Two regression coefficients reported in Borjas (1994) lie at the core of the ALV reevaluation.[1] The first coefficient comes from a regression of third-generation persons' educational attainment on the 1910 literacy rate of the immigrant group that represents each person's ethnic ancestry. The regression coefficient is 0.025 (standard error, s.e. = 0.018). The second coefficient comes from a regression of the mean log occupational wage of third-generation workers on the 1910 mean log wage of the immigrant group representing the worker's ethnic ancestry.[2] The coefficient is 0.216 (s.e. = 0.108). In my original discussion I stressed the coefficient obtained from the wage regression. Strangely, the ALV critique focuses almost exclusively on the coefficient obtained from the education regression, which was not statistically significant in my paper.

The ALV critique makes three points:

First, Borjas (1994) ignored the mixed ethnic backgrounds of many third-generation Americans. In my empirical analysis, which used the 1977–1989 GSS cross-sections, I defined ethnicity using the "main" ethnic background identified by the respondent. Beginning with the 1986 survey, the GSS has allowed respondents to identify up to three distinct ethnic ancestries. ALV use this information to calculate an "average ethnic score" for the respondent's ancestors. For example, if a person reports both Irish and Italian ancestry, the average ethnic score would be the mean of these two groups' literacy rates.

Second, ALV show that the correlation between the skills of the first and the third generations becomes much weaker if some ethnic groups are excluded from the regression. In particular, much of the intergenerational linkage reported in my paper is generated by the presence of persons of Mexican ancestry in the sample. ALV propose various reasons why these persons should be excluded.

Third, political upheavals redrew the map of Europe during the twentieth century; thus it was difficult to match the GSS respondents' self-reported ethnic origin with the actual ethnic composition of the European political units before World War I. ALV's solution to this problem is to exclude more ethnic groups, specifically those originating from Austria, Hungary, Poland, and Yugoslavia.

ALV's own evidence indicates that only the second of these criticisms (the inclusion of Mexicans in the analysis) matters empirically. ALV's table 2 shows that the results are not affected when one changes the definition of ethnicity from the respondent's main ethnic background to an "average" ethnicity (compare the third and the fourth panels of

*George J. Borjas, Pforzheimer Professor of Public Policy, Kennedy School of Government, Harvard University; and research associate, National Bureau of Economic Research. Direct correspondence to George J. Borjas, Harvard University, John F. Kennedy School of Government, 79 John F. Kennedy Street, Cambridge, MA 02138; E-mail: Gborjas@harvard.edu.

[1] The coefficients appear in the second row of both panels of table 8 (Borjas 1994). In preparing this response, I discovered that the information reported in the bottom panel is mislabeled, although the discussion in the text defines precisely what the coefficients measure. The column headings for the bottom panel should read "Parental log wage," "Mean log wage in parents' generation," and "Log wage of 1910 immigrants." ALV do not mention the labeling problem, but their discussion indicates that they interpreted my results correctly.

[2] Neither the 1910 census nor the GSS reports a worker's actual earnings; instead both data sets report the worker's occupation. I calculated a worker's log occupational wage by assigning each worker the mean value of the log wage in the occupation in which he or she was employed. The mean log wage for each occupation was obtained from other data sources, such as the 1970 census.

the table). Moreover, the last two columns of the table show that the results are not affected when ALV exclude additional ethnic groups to "control" for the confusion created by the redrawing of Europe's map.

For my response, I obtained the GSS data directly from ALV. As a result, I use exactly the same GSS data extract, as well as the same measures of the immigrant group's human capital as of 1910 (reported in Borjas 1994: table 3).

Before proceeding to my appraisal of the ALV study, I must derive the statistical model underlying the calculations to illustrate some statistical problems and conceptual flaws that mar the ALV critique. The regression model estimated by ALV is given by

$$y_{ij} = \boldsymbol{Z}_{ij}\alpha + \delta x_j + \varepsilon_{ij}, \qquad (1)$$

where y_{ij} is some measure of the skills of person i in ethnic group j in the third generation (e.g., educational attainment); $\boldsymbol{Z}_{ij}$ is a vector of standardizing characteristics; and x_j is the measure of the average skills for ethnic group j in the first generation (e.g., the literacy rate). It is instructive to provide a more detailed derivation of this "mixed" regression model (in the sense that it uses individual-level data on the left-hand side and group-level data on the right-hand side). In particular, consider a regression model that attempts to estimate the adjusted skill differentials across ethnic groups in the third generation. This regression would be given by

$$y_{ij} = \boldsymbol{Z}_{ij}\beta + \lambda_j + \mu_{ij}, \qquad (2)$$

where λ_j is a fixed effect for the ethnic group. These fixed effects measure the adjusted differences in skills across ethnic groups in the third generation. From the perspective of understanding the intergenerational persistence of ethnic skill differentials, the regression of interest is provided by linking these third-generation fixed effects to the mean skills of the immigrant group, or

$$\lambda_j = \gamma + \delta x_j + v_j. \qquad (3)$$

One can see that Eq. (1), estimated by ALV in their table 2, may be obtained by substituting Eq. (3) into Eq. (2). This substitution reveals that the error term in the ALV model is $\varepsilon_{ij} = v_j + \mu_{ij}$, so that it contains both a group-specific (v_j) and a person-specific (μ_{ij}) component.[3] The group-specific part of the error term implies that there is a correlation in the residual among observations that belong to the same ethnic group in the mixed regression. For example, if it is believed that the U.S. labor market penalizes ethnic group j for having a particular skin color or speaking a particular language, the random variable v_j might be negative; this would impart a correlation in the error term among all type j persons in the GSS sample.

As argued in my original paper, one should interpret the evidence reported in table 2 of ALV's critique in terms of this two-stage framework. The correct number of *independent* observations on ethnic groups' skills is not given by the number of respondents in the GSS; there is really only *one* observation per ethnic group. As a result, the two-stage framework raises serious conceptual questions about the cavalier way in which ALV drop ethnic groups from the GSS sample. Moreover, the correlation in the residual across observations implies that the standard error of δ estimated by an ordinary least squares regression of the mixed model, and reported throughout ALV's table 2, is incorrect (see Moulton 1986).

The first row of Table 1 displays my reestimation of the ALV model. To simplify the discussion, I replicate the final specification that ALV use in their table 2 (fourth row). This specification accounts for the GSS respondents' mixed ethnic background, so that the regression employs the 1986–1994 GSS sample and uses "averaged ethnic scores" as the independent variable.[4] Not surprisingly, I can replicate their work completely because we are all using the same data sources. In the second row of my table I reestimate the ALV regression, this time correcting for the error structure of the data.[5] The correct standard errors are often twice as large as those reported by ALV. Ironically, more careful attention to the statistical methodology (and to the conceptual model underlying the regression) would have strengthened the general point made by ALV: the link between third-generation educational attainment and first-generation literacy rates is statistically insignificant.

Just before reaching the conclusion to their critique, ALV insert a peculiar paragraph stating that "to be fair," they could have obtained different results if they had used a different measure of the immigrant group's human capital or economic status in 1910. In my original paper, as noted above, I used two alternative measures of the ethnic group's economic status: the literacy rate and the mean wage. It turns out that the nature of the paper written by ALV would have changed substantially if they had chosen to stress regressions of the third-generation educational attainment on the mean log wage of the 1910 immigrant group.[6] The third row of Table 1 presents the relevant coefficients.

In column 1, the coefficient is 3.99 (s.e. = 1.53). Excluding Mexicans and two other, smaller groups reduces the estimated coefficient to 1.82, but the coefficient is still somewhat larger than its standard error. Almost all of the change in the magnitude of the coefficient results from deleting persons of Mexican ancestry from the GSS sample.[7] The table also shows that the coefficient would have remained in the 4.0 range (with relatively low standard errors) if, instead of excluding Mexicans, I had excluded respondents who origi-

[3] The random variables v_j and μ_{ij} are typically assumed to be uncorrelated.

[4] One could quibble over whether the "average ethnic score" is conceptually preferable to using the main ethnicity identified by the respondent. Both approaches impose arbitrary weights on the data. There is little need to assess the relative merits of the two definitions because they lead to similar results.

[5] I calculate the standard errors using STATA's cluster option in linear regression models.

[6] Following ALV, I defined the average ethnic score in the wage regressions as the simple average of the occupational wage of the various immigrant groups identified by the third-generation workers.

[7] The coefficient is 1.41 (s.e. = 1.18) if only persons of Mexican ancestry are excluded.

TABLE 1. REAPPRAISAL OF EVIDENCE LINKING THE FIRST AND THIRD GENERATIONS: POOLED 1986–1994 GSS CROSS-SECTIONS, USING AVERAGE ETHNIC SCORES

	Regression					
Specification	(1)	(2)	(3)	(4)	(5)	(6)
Education Regressed on 1910 Literacy Rates						
Replicating Alba, Lutz, and Vesselinov	0.017	–0.004	0.022	0.001	0.004	0.009
	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)
	[2,153]	[2,072]	[1,785]	[1,706]	[1,898]	[1,584]
Corrected standard errors	0.017	–0.004	0.022	0.001	0.004	0.009
	(0.013)	(0.010)	(0.013)	(0.010)	(0.006)	(0.004)
	[2,153]	[2,072]	[1,785]	[1,706]	[1,898]	[1,584]
Education Regressed on 1910 Log Wage Rate	3.989	1.822	4.460	2.718	0.503	1.346
	(1.532)	(1.262)	(1.408)	(1.230)	(1.155)	(1.146)
	[2,153]	[2,072]	[1,785]	[1,706]	[1,898]	[1,584]
Log Wage Regressed on 1910 Log Wage Rate	0.269	0.147	0.328	0.261	0.105	0.218
	(0.118)	(0.119)	(0.112)	(0.157)	(0.125)	(0.163)
	[2,081]	[2,004]	[1,723]	[1,648]	[1,834]	[1,529]
Excludes China, Japan, and Mexico	No	Yes	No	Yes	Yes	Yes
Excludes Austria, Hungary, Poland, and Yugoslavia	No	No	Yes	Yes	No	Yes
Excludes Romania and Russia	No	No	No	No	Yes	Yes

Notes: Standard errors are shown in parentheses; sample size of GSS data are shown in brackets. All standard errors except those reported in the first row are corrected for potential correlations in the error term within ethnic groups. All regressions include the GSS respondent's age, gender, region of residence, and metropolitan residence, as well as fixed effects indicating the GSS cross-section from which the observation is drawn. The average ethnic score used in the first and second rows is the simple average of the literacy rate of the ethnic groups that make up a person's ethnic ancestry. The average ethnic score used in the third and fourth rows is the simple average of the respective ethnic groups' mean log wage.

nated in countries such as Austria, Hungary, Poland, and Yugoslavia (presumably to control for the changes in the European map). Finally, column 4 reports the results based on the preferred ALV specification, which excludes persons from both sets of countries. The coefficient is now 2.72 *and* is statistically significant (s.e. = 1.23). In contrast to the evidence provided by the regressions of third-generation education on literacy rates, the regressions of third-generation education on immigrants' wages provide limited support for an inference that, even in a selected set of ethnic groups of European origin, "ethnicity matters, and it matters for a very long time" (Borjas 1994:572).

Even if one accepts the ALV rationale for excluding various ethnic groups from the analysis (which I dispute below), the quantitative impact of the intergenerational correlation is not small. Consider, for example, the long-run implications of a .20 difference in mean log wages between two immigrant groups in 1910, equivalent to the "wage distance" between England and Italy.[8] Even after a century, the coefficient reported in column 4 implies that this initial wage differential is associated with a 0.54-year difference in the grandchildren's educational attainment. If the rate of return to education were on the order of 10%, the initial 20% wage gap would imply a 5% wage gap in the third generation. Moreover, this finding is statistically significant.

ALV are obviously aware of these results. Although they do not report this evidence fully in their paper, they observe that the statistical significance would vanish if, in addition to excluding the Mexicans, the Chinese, and the Japanese, one also excluded Jews from the sample (specifically persons of Romanian or Russian ancestry). In fact, column 5 of Table 1 shows that the regression coefficient falls to 0.50 (s.e. = 1.16) when I omit these two additional groups. What is ALV's rationale for excluding Romanian and Russian Jews? Doing so, they say, shows that the significant positive correlation "is due chiefly to the extraordinary educational attainments...(of) groups for which one would predict a low educational trajectory because of their peasant origins in Europe" (p. 355). In other words, the intergenerational correlation disappears when one filters the data to exclude groups that have done extraordinarily well.

To gain a sense of the magnitude of the ethnic exclusions imposed by ALV, consider the following exercise. The GSS

[8] The mean log wage in 1910 was 6.43 for English immigrants and 6.23 for Italian immigrants.

sample contains 473 unique combinations of ethnic ancestries (e.g., Italian, Mexican-Irish, German-Scottish-Russian). If one makes the exclusions that ALV make in their table 2 (specifically Austria, China, Hungary, Japan, Mexico, Poland, and Yugoslavia), the remaining sample contains 317 such unique combinations. The ALV exclusions therefore remove 33% of the independent observations that potentially could have been used in the second-stage regression.

The statistical approach pursued by ALV offers a disturbing lesson for social science research. At the very least, it ignores 30 years of insights from economics on the statistical problems associated with sample selection.[9] At worst, it conveys the message that one can achieve a desired result (i.e., that δ is zero) by identifying those observations that generate significant coefficients and systematically removing them from the sample. Consider, for instance, the contrasting rationales used by ALV for excluding Mexicans and Jews from the third-generation sample. On the one hand, Mexicans are excluded "because they suffered more severe liabilities.... Quite obviously these liabilities may have interfered with any continuous process of adjustment and socioeconomic upgrading" (p. 350). On the other hand, Jews are excluded because they have more education than would have been expected in view of their "peasant origins." The ALV statistical approach thus can be summarized as follows: Mexicans should be left out because they are doing too poorly, and Jews should be left out because they are doing too well.

It is worth interpreting the ALV approach in terms of the model presented above. Persons of Mexican ancestry are excluded because the conditional expectation E(v_j | x_j, Mexican ancestry) is much too negative. This disadvantage presumably implies that the Mexican experience cannot provide any valuable information about the intergenerational mobility experienced by the "typical" European group. Similarly, Jews are excluded because E(v_j | x_j, Jewish ancestry) is too positive; this advantage presumably implies that the Jewish experience cannot provide valuable information about the intergenerational mobility experienced by the "typical" European group.

What exactly is the statistical problem that is solved by removing these ethnic groups from the sample? Certainly it cannot be that the inclusion of these groups would imply that the random variable v has a nonzero mean. After all, a nonzero mean of the error term would affect only the constant term, and not the parameter of interest (δ). Similarly, it cannot be that the inclusion of these groups would bias the results because the random variable v would then be heteroscedastic: that is, v would not have a constant variance because of the very different experiences of Mexicans and of Jews in the United States. It is well known that heteroscedasticity does not bias the coefficients, and its impact on the standard error of the parameter δ could easily be addressed by estimating the model using generalized least squares. Conceivably one could argue that perhaps the unobserved random variable v_j and the initial conditions (x_j) are correlated, thus biasing estimates of the parameter δ. Yet it is far from clear how the exclusion of these specific groups solves the endogeneity problem.

These technical concerns could be dismissed by arguing that the parameter of interest is the intergenerational correlation experienced by European groups, and that the parameter δ for Europeans might differ from the δ for other groups. But then why omit the Russians and the Romanians? Or, perhaps more perversely, why stop there? After all, other European ethnic groups surely have similar idiosyncratic histories. Two other such groups can be identified quickly: the Germans and the Italians. The assimilation experiences of these two large groups surely were affected by the fact that Germany and Italy were on the wrong side of wars fought by the United States in the twentieth century. One should not dismiss the importance of armed conflict in determining assimilation rates. Conzen (1980:423) reports that "by summer 1918 about half of the [U.S.] states had restricted or eliminated German-language instruction, and several had curtailed freedom to speak German in public.... The total number of German-language publications declined from 554 in 1910 to 234 in 1920."

I am *not* arguing that one should continue omitting groups because of these unique historical circumstances. These circumstances are precisely the factors that shape the distribution of the random variable v_j in the second-stage regression. In some cases, perhaps because of the color of their skin, some groups will have negative vs; thus, despite similar starting conditions, these disadvantaged groups will have experienced less upward mobility during the twentieth century. Other groups, perhaps because they settled in booming areas, will have positive vs. Such idiosyncratic events do not justify the exclusion of these groups from the statistical analysis.

ALV devote their entire paper to the sensitivity of one of the two coefficients that I reported in my original work; they do not examine the robustness of the other. The "missing" coefficient comes from a regression of the log occupational wage of third-generation workers on the mean log wage of the immigrant group in 1910. ALV claim that they "could not easily reproduce [the analysis of third-generation wages] because the variable is a construction based on the average wages in the 1970 census for each occupation" (p. 353). To illustrate the sensitivity of the intergenerational wage correlation to the "corrections" proposed by ALV, I used the sample of workers age 25 to 64 in the 1970 and 1980 Public Use Microdata Samples (PUMS) of the census to calculate the mean log wage rate in each occupation. I then used these data to assign each worker in the GSS a log occupational wage.[10]

[9] Heckman (1979) gives the classic statement of the selection problem.

[10] Until 1990 the GSS used the 1970 census codes to classify the respondents' occupation. In 1988 the GSS began to use the 1980 census codes. For the years between 1988 and 1990, the GSS reports the worker's occupation using both codes. I used the 1970 census data to impute the occupational wage of workers in cross-sections before 1988, and the 1980 census data to impute the occupational wage beginning with the 1988 data.

The fourth row of Table 1 shows the coefficients from regressions in which the dependent variable is the log wage of the third-generation worker and the independent variable is the mean log wage of the relevant ethnic ancestors as of 1910. I find a positive, and sometimes significant, correlation between these two variables. In the absence of any of the sample exclusions made by ALV, the coefficient is 0.269 (s.e. = 0.118). If I exclude from the regression the same two sets of ethnic groups excluded by ALV in their table 2, the coefficient is 0.261 (s.e. = 0.157). If, in addition, Romanian and Russian Jews are excluded from the analysis, the coefficient declines to 0.218 (s.e. = 0.163).

What should one conclude from my Table 1? In the end, the ALV critique hinges on two crucial points. First, it seems that only one possible specification of the intergenerational mobility regression is valid in their perspective, namely the regression of third-generation educational attainment on the first generation's literacy rates. Yet the intergenerational transmission of skills need not be observed solely through educational attainment; it can also be observed through wages. Although not estimated precisely, the coefficients in the wage regressions suggest some link between the wages of the first and the third generations, even after exclusion of various ethnic groups that account for much of the variation in the data.

In fact, one could make a strong case that the wage regressions capture the parameter of interest far more accurately than data on educational attainment and literacy rates. Education explains only a small part of wage variation in the population, and much of what one would regard as socioeconomic success or failure would be lost by focusing exclusively on educational attainment. In the present context, the intergenerational correlation between education and literacy rates would not capture the possible upward mobility experienced by some relatively low-skill groups in which many immigrants became shopkeepers and invested heavily in their children's schooling and other forms of human capital. In contrast, the wage data would measure the relative success of these groups and of their children. Stated bluntly, there is no reason to dismiss, as do ALV, the correlations between third-generation socioeconomic achievement and first-generation wages.[11]

Second, and perhaps more important, I am disconcerted by the ALV approach of systematically excluding those ethnic groups that do not seem to fit their preconceived notion of a "normal" assimilation pattern. Their paper and my response surely have shown that by systematically excluding ethnic groups from the analysis (particularly Mexicans and Jews), one can reduce to zero any estimated correlation between immigrants's skills and those of their grandchildren. Does that approach provide a valuable road map for future research? I doubt it.

REFERENCES

Alba, R., A. Lutz, and E. Vesselinov. 2001. "How Enduring Were the Inequalities Among European Immigrant Groups in the United States?" *Demography* 38:349–56.

Borjas, G. 1994. "Long-Run Convergence of Ethnic Skill Differentials: The Children and Grandchildren of the Great Migration." *Industrial and Labor Relations Review* 47:553–73.

Conzen, K. 1980. "Germans." Pp. 405–25 in *Harvard Encyclopedia of American Ethnic Groups,* edited by S. Thernstrom. Cambridge, MA: Harvard University Press.

Heckman, J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47:153–61.

Moulton, B. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics* 32:385–97.

Solon, G. 1999. "Intergenerational Mobility in the Labor Market." Pp. 1761–800 in *Handbook of Labor Economics,* Vol. 3A, edited by O. Ashenfelter and D. Card. Amsterdam: Elsevier.

[11] Many of the recent studies on intergenerational transmission use wage data in the analysis. Solon (1999) surveys this literature.