Linking ACS and IRS Data to Assess Educational Attainment by Family Income

Leah Clark, Jennifer Ortman, Nikolas Pharris-Ciurej & John Voorheis
U.S. Census Bureau

Any views expressed are those of the authors and not those of the U.S. Census Bureau. The Census Bureau's Disclosure Review Board and Disclosure Avoidance Officers have reviewed this data product for unauthorized disclosure of confidential information and have approved the disclosure avoidance practices applied to this release. DRB approval number CBDRB-FY2020-CES010-008
Tracking educational attainment by childhood family income

- Cross-sectional surveys like the American Community Survey (ACS) offer measures of educational attainment on a national scale
  - Pros: Large sample size, collected annually
  - Con: Lacks reliable information about childhood family income
- Studying this topic with other data sources brings up other challenges
  - Longitudinal surveys have information about childhood family income but smaller sample size and fewer cohorts
    - For example, Bailey & Dynarski (2010) use National Longitudinal Surveys of Youth data to assess college attendance and completion, but are limited to two cohorts
  - Administrative records offer annual population-level statistics, but are limited in scope
    - For example, Chetty et al. (2014) use tax records to study college attendance, but cannot assess college completion
Our contribution

• We link ACS data to Internal Revenue Service (IRS) form I-1040 data to bring a measure of childhood family income into the ACS
  • Making a cross-sectional survey longitudinal by leveraging administrative data (Dynarski 2014)
• We offer a comprehensive picture of high school completion, college attendance, and college completion for the same respondents
  • We focus on cohorts born from 1983-1991
    • Potential to expand coverage going forward
• We produce reliable, annual statistics on educational attainment by childhood family income
  • Paired with detailed demographic and geographic characteristics, this represents a major step forward for measuring inequality in educational attainment
Trends in high school completion

Related literature
• Heckman & LaFontaine (2010) document a stagnant high school graduation rate between 1970 and 2000
  • Stress difference between graduation (with a regular diploma) and completion (with a GED or alternative diploma)
• Murnane (2013) finds growth in high school graduation rates since 2000, especially for Black and Hispanic students

Our findings
• Upward trend in high school completion driven by gains for low-income young people
  • Narrows income gap by more than one-fourth
  • For low-income young people, Black-White and Hispanic-White gaps narrow, but AIAN-White gaps grow
Trends in college attendance and completion

Related literature

• Bailey & Dynarski (2011) find growth in college attendance and completion for cohorts born 1961-64 and 1979-82
  • Growth favored young people from high income families, widening income gaps
• Chetty et al. (2014) find slight reduction in college attendance gradient (gap) for young people born in the 1980s

Our findings

• Large income gaps in college attendance and completion
  • Persist within racial subgroups
• Income gaps in college attendance narrow very slightly for cohorts born 1983-1991
• Income gaps in college completion widen as low- and middle-income rates stagnate, but high-income rates continue to rise
Defining educational outcomes

- Derived from ACS question concerning *highest* level of educational attainment
- High school completion includes regular high school diploma, GED/alternative credential attainment or higher
- College attendance includes “some college credit” or higher
- College completion is defined as a bachelor’s degree or higher
  - Measures that include associate’s degrees are also an option
1. Locate ACS respondents as dependents in IRS form I-1040 data
2. Pull household adjusted gross income during the years they turned 15, 16, or 17, and adjust for inflation
3. Define childhood family income as the average of those three years, ignoring missing or negative values
4. Rank respondents by childhood family income within each birth cohort and split into three equal-sized groups (high, middle, low)
New linked dataset

ACS 2006-2017
  - Surveyed at ages 24-26
- Variables:
  - Educational attainment
  - Race/ethnicity
  - Sex
- Excludes noncitizens

Administrative Records
- IRS form I-1040
  - 1998-2014
  - Variable: Childhood family income
- Census Numident
  - Social Security Administration records
  - Variable: Date of birth

= Over 1 million linked respondents
- Childhood family income measures for 87% of the population
- Sample weights rescaled by inverse probability of having a measure of childhood family income (details)
Educational attainment by age

What is the educational attainment of young people born between 1983 and 1991?

Conditional rates allow us to isolate the inequality that arises at each education level.

- 72% of high school completers attend college.
- 44% of college attenders complete college.

How does educational attainment vary by family income?

High-income young people are 1.2x as likely to finish high school as low-income young people.

High-income high school completers are 1.5x as likely to attend college as low-income high school completers.

High-income college attenders are 2.3x as likely to finish college as low-income college attenders.


[Appendix: outcomes by noniles]
Race/ethnicity

- We use the following race/ethnicity categories:
  - Hispanic (of any race), and
  - Non-Hispanic:
    - White
    - Black
    - Asian
    - American Indian/Alaskan Native (AIAN)
    - Native Hawaiian/Pacific Islander (NHPI)
    - Other

[Appendix: income categories by race]

High school completion by income & race


[Appendix: Outcomes by race and sex]
Conditional college attendance by income & race

Conditional college completion by income & race

Trends in educational attainment

• Rather than pool over birth years, we can plot attainment for each year and income group

• To test changes in income gaps, we regress each outcome on year-by-income group dummies, omitting the first year and middle-income respondents
  • A coefficient of .04 for low-income young people in 1991 for high school completion means that…
    • the low-income high school completion rate gained 4 percentage points
    • since 1983
    • relative to the middle-income rate
Low-income young people made gains in high school completion

[Appendix: terminal GED/alt diplomas do not change trends]
Small low-income gains in college attendance

[Appendix: age of measurement matters]
The income gap widened in college completion

Unequal gains in high school completion for low-income young people with different racial/ethnic backgrounds


- For low-income American Indian/Alaskan Native respondents, however, growth in high school completion did not keep up with their White counterparts.
  - The low-income AIAN-White high school completion gap grew.

[Appendix: college outcomes]

Next steps

• State-level estimates
• Analysis for additional demographic groups of interest
  • Foreign-born young people; foreign-born parents
Conclusion

• Linking longitudinal/cross-generational measures from administrative data to cross-sectional surveys opens many new analysis opportunities

• High school completion rates are rising
  • Income gaps in high school completion narrowed somewhat for cohorts born in the 1980s, and are likely continuing to converge
  • Gains are not evenly distributed across subgroups

• College attendance and completion rates are stable
  • Income gaps in college attendance and completion are large and persistent

• Income gaps vary by race and gender subgroups, but they are consistently large within race and gender subgroups

• American Indian and Alaskan Native young people experience low levels of educational attainment, and are not showing signs of convergence with other subgroups
Questions?

Leah.R.Clark@census.gov
Appendix
Adjusting weights to account for data linkage

• Some ACS respondents cannot be linked to their child family income
  • Census uses personally identifiable information (PII) to link survey/administrative records to a protected identification key (PIK) that is used for linking across sources
    • Sometimes records cannot be assigned a PIK
  • Some parents never filed a tax return while respondents were ages 15-17, did not claim them as a dependent, or reported negative income

• Sample weights need to be adjusted to account for nonrandom probability of being able to derive child family income
  • Logistic regression of the probability of linking an ACS respondent to their child family income on covariates including respondent’s relation to householder (e.g., spouse, son/daughter), sex, age, race, naturalization status, state of residence
    → Estimate predicted probability of linking to child family income
  • Sample weights are scaled by the inverse of that predicted probability
Race/ethnicity and income

- Income categories are defined on full sample, and are not redefined within subgroups
  - Young people of color – especially Black, Hispanic, and AIAN young people – have disproportionately low childhood family income

- Young people of color are also more likely to not be linked to childhood family income
  - Rescaling sample weights to account for nonrandom linkage offers an imperfect remedy

## Educational attainment by income, race, and gender

<table>
<thead>
<tr>
<th></th>
<th>High school completion</th>
<th>Conditional college attendance</th>
<th>Conditional college completion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>Middle</td>
<td>High</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td>0.84</td>
<td>0.93</td>
<td>0.98</td>
</tr>
<tr>
<td>White Female</td>
<td>0.88</td>
<td>0.95</td>
<td>0.98</td>
</tr>
<tr>
<td>Hispanic Female</td>
<td>0.85</td>
<td>0.92</td>
<td>0.96</td>
</tr>
<tr>
<td>Black Female</td>
<td>0.87</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>Asian Female</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>AIAN Female</td>
<td>0.80</td>
<td>0.91</td>
<td>0.94</td>
</tr>
<tr>
<td>NHPI Female</td>
<td>0.89</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>Other Female</td>
<td>0.90</td>
<td>0.95</td>
<td>0.99</td>
</tr>
<tr>
<td>White Male</td>
<td>0.84</td>
<td>0.93</td>
<td>0.98</td>
</tr>
<tr>
<td>Hispanic Male</td>
<td>0.80</td>
<td>0.88</td>
<td>0.94</td>
</tr>
<tr>
<td>Black Male</td>
<td>0.78</td>
<td>0.90</td>
<td>0.95</td>
</tr>
<tr>
<td>Asian Male</td>
<td>0.94</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>AIAN Male</td>
<td>0.76</td>
<td>0.88</td>
<td>0.91</td>
</tr>
<tr>
<td>NHPI Male</td>
<td>0.88</td>
<td>0.91</td>
<td>0.97</td>
</tr>
<tr>
<td>Other Male</td>
<td>0.87</td>
<td>0.91</td>
<td>0.96</td>
</tr>
</tbody>
</table>

## Income gaps by race and gender

<table>
<thead>
<tr>
<th></th>
<th>High school completion</th>
<th>Conditional college attendance</th>
<th>Conditional college completion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute High-Low</td>
<td>Relative High/Low</td>
<td>Absolute High-Low</td>
</tr>
<tr>
<td>All</td>
<td>0.13</td>
<td>1.16</td>
<td>0.27</td>
</tr>
<tr>
<td>White Female</td>
<td>0.10</td>
<td>1.12</td>
<td>0.25</td>
</tr>
<tr>
<td>Hispanic Female</td>
<td>0.11</td>
<td>1.13</td>
<td>0.20</td>
</tr>
<tr>
<td>Black Female</td>
<td>0.10</td>
<td>1.11</td>
<td>0.22</td>
</tr>
<tr>
<td>Asian Female</td>
<td>0.03</td>
<td>1.03</td>
<td>0.08</td>
</tr>
<tr>
<td>AIAN Female</td>
<td><strong>0.14</strong></td>
<td><strong>1.17</strong></td>
<td>0.19</td>
</tr>
<tr>
<td>NHPI Female</td>
<td>0.06</td>
<td>1.07</td>
<td>0.24</td>
</tr>
<tr>
<td>Other Female</td>
<td>0.09</td>
<td>1.10</td>
<td>0.16</td>
</tr>
<tr>
<td>White Male</td>
<td><strong>0.14</strong></td>
<td><strong>1.16</strong></td>
<td><strong>0.32</strong></td>
</tr>
<tr>
<td>Hispanic Male</td>
<td><strong>0.15</strong></td>
<td><strong>1.18</strong></td>
<td>0.24</td>
</tr>
<tr>
<td>Black Male</td>
<td><strong>0.17</strong></td>
<td><strong>1.22</strong></td>
<td>0.26</td>
</tr>
<tr>
<td>Asian Male</td>
<td>0.05</td>
<td>1.05</td>
<td>0.10</td>
</tr>
<tr>
<td>AIAN Male</td>
<td><strong>0.15</strong></td>
<td><strong>1.19</strong></td>
<td>0.25</td>
</tr>
<tr>
<td>NHPI Male</td>
<td>0.09</td>
<td>1.10</td>
<td>0.23</td>
</tr>
<tr>
<td>Other Male</td>
<td>0.09</td>
<td>1.10</td>
<td>0.22</td>
</tr>
</tbody>
</table>

**Bold** values indicate larger income gaps than is true for the full population.

Family income and educational attainment

Does this mean more young people are graduating from high school?

- Heckman & LaFontaine (2010)
  - High school completion rate overestimates graduation rate
  - Due primarily to GED attainment
    - GEDs are not economically equivalent to diplomas
- Self-reported attainment may differ from administrative attainment
  - New alternative diplomas
- We can remove *terminal* GED and alternative diplomas from our high school completion measure
  - Still find uptick

Note: Dashed lines show high school completion rate when individuals who report a GED or alternative diploma as their highest degree are not counted as completers.

Excluding terminal GED/alternative diplomas yields parallel trends, and a larger income gap

Age matters in measuring college attendance rates

- 20-22-year-olds born in the late 80s show large increases in college attendance rates
  - But the same cohorts surveyed at ages 24-26 show much smaller gains
- Possible explanation: around the recession, young people chose not to delay college
  - Intertemporal substitution, not real growth in college attendance

Relative changes in college outcomes for low-income young people by race/ethnicity

College attendance
(24-26-year-old high school completers)

Change since 1983-85 relative to White

College completion
(24-26-year-old college attendees)

Change since 1983-85 relative to White