Journal of Educational Psychology
2013, Vol. 105, No. 1, 176–198

© 2012 American Psychological Association
0022-0663/13/$12.00 DOI: 10.1037/a0029481

When Less Is More: Effects of Grade Skipping on Adult STEM Productivity Among Mathematically Precocious Adolescents

Gregory Park, David Lubinski, and Camilla P. Benbow
Vanderbilt University

Using data from a 40-year longitudinal study, the authors examined 3 related hypotheses about the effects of grade skipping on future educational and occupational outcomes in science, technology, engineering, and mathematics (STEM). From a combined sample of 3,467 mathematically precocious students (top 1%), a combination of exact and propensity score matching was used to create balanced comparison groups of 363 grade skippers and 657 matched controls. Results suggest that grade skippers (a) were more likely to pursue advanced degrees in STEM and author peer-reviewed publications in STEM, (b) earned their degrees and authored their 1st publication earlier, and (c) accrued more total citations and highly cited publications by age 50 years. These patterns were consistent among male participants but less so among female participants (who had a greater tendency to pursue advanced degrees in medicine or law). Findings suggest that grade skipping may enhance STEM accomplishments among the mathematically talented.

Keywords: educational acceleration, gifted, math/science talent, longitudinal analysis, propensity score matching

Grade-based acceleration, or grade skipping, is an educational intervention targeted at intellectually precocious students with the goal of allowing such students to experience more developmentally appropriate content by skipping over what they already know or can easily and rapidly assimilate (Colangelo, Assouline, & Lupkowski-Shoplik, 2004; Stanley, 2000). Empirical research on the short-term effects of specific forms of grade-based acceleration, such as early entrance to kindergarten and early entrance to college, has been supportive (Kulik & Kulik, 1984; Rogers, 2007; Stanley, 1973; Swiatek & Benbow, 1991). On the other hand, much less is known about its long-term effects (e.g., 20 years later).

Despite the scarcity of empirical studies, interest in long-term effects of acceleration, broadly defined, and grade skipping in particular has a long history, spiking in successive postwar periods in the early and mid-20th century (Hobbs, 1951; Paterson, 1957; Seashore, 1922; Super & Bachrach, 1957). A recurring idea is that many intellectually talented students could benefit from increasing the rate at which they move through the educational system, and in turn, the arts, sciences, and technological fields could reap benefits as well (Pressey, 1955; Terman, 1954). Pressey (1946b) argued that grade-based acceleration had the potential to save valuable time during a critical period in a precocious individual's development and offered a theory of how acceleration may increase overall career productivity. Individuals, Pressey (1946b) suggested, have a "prime" in early adulthood during which the probability of illness and death are at a low level, and positive attributes such as strength, quickness of body and mind, and vigor of interests are at their peak.[1] During this critical period, those students pursuing advanced training in scientific and technical fields are often bogged down in training rather than actively producing, and this may "curtail maximum fruitfulness of a professional career" (Pressey, 1946a, p. 324). Likewise, Terman (1954) stressed the need to capitalize on this developmental prime by training those with high potential "before too many of his most creative years have been passed" (Terman, 1954, p. 226). It was reasoned that if the brightest students could advance more quickly through the educational system, they would lose little, if anything, but potentially gain intellectual development, interpersonal maturity, and, most importantly, time. This could lead to higher levels of career productivity and creative accomplishments.

This article was published Online First August 13, 2012.

Gregory Park, David Lubinski, and Camilla P. Benbow, Department of Psychology and Human Development, Vanderbilt University.

This study was based on a dissertation submitted to Vanderbilt University by Gregory Park in partial fulfillment of the doctor of philosophy degree. Support for this study was provided by a research and training grant from the Templeton Foundation and by National Institute of Child Health and Development Grant P30 HD 15051 to the Vanderbilt Kennedy Center for Research on Human Development. Earlier versions of this article benefited from comments by Kimberley Ferriman Robertson, Stijn Smeets, James H. Steiger, Andrew J. Tomarken, and Jonathan Wai.

Correspondence concerning this article should be addressed to Gregory Park, who is now at the Positive Psychology Center, University of Pennsylvania, 3701 Market Street, Suite 219, Philadelphia, PA 19104; or to David Lubinski or Camilla P. Benbow, Department of Psychology and Human Development, Vanderbilt University, 0552 GPC, 230 Appleton Place, Nashville, TN 37203. E-mail: gregoryjpark@gmail.com, david.lubinski@vanderbilt.edu, or camilla.benbow@vanderbilt.edu

[1] An updated interpretation of the time-saving theory may add competing interests (e.g., work vs. family), work preferences (e.g., overtime vs. full-time vs. part-time), and other responsibilities to the list of "threats" looming in early adulthood; these factors are likely to influence individuals' career choices differently throughout early adult development (Ferriman et al., 2009).

We refer to this theory as the *time-saving theory* (Pressey, 1946b). According to this theory, grade-based acceleration can directly affect a precocious individual's career trajectory in two ways: by increasing the likelihood that they pursue advanced degrees and training and by allowing them to finish this training earlier. First, if precocious individuals finish high school and undergraduate programs earlier than usual, they will be more likely to reinvest this time in themselves through additional education and training. Second, those who do reinvest this time will finish developing expertise earlier than their retained (nonskipping) peers, and this will allow an earlier career onset. The relationship between age at career onset and adult productivity, particularly in science, technology, engineering, and mathematics (STEM) fields, has been the focus of several researchers throughout the last century (Dennis, 1956; Lehman, 1946, 1953; Simonton, 1988, 1997; Zuckerman, 1977), and a consistent finding is that earlier career onset is related to greater productivity and accomplishments over the course of a career. All other things being equal, an earlier career start from acceleration will allow an individual to devote more time in early adulthood to creative production, and this will result in an increased level of accomplishment over the course of one's career.

Within the last decade, the rapid increase in globalization led to a resurgence of interest in boosting productivity and enhancing national competitiveness, most notably in the fields of science, technology, engineering, and mathematics (STEM; e.g., Domestic Policy Council & Office of Science and Technology Policy, 2006; Friedman, 2005; National Science Board, 2010a). Scientific and technological innovation are recognized drivers of national economic growth and quality of life improvement, and the identification and development of STEM talent is a regional and national concern. The time-saving theory suggests that the increased application of grade-based acceleration may be one means of addressing this issue. However, the compelling data to support this theory empirically have been scarce for several reasons. First, research on long-term effects on career productivity requires longitudinal data spanning several decades. Second, methodological issues common to educational research, such as selection bias or imbalance, compromise causal inferences about effects even when longitudinal data are available. Third, studies of the long-term effects of grade skipping on rare STEM outcomes require a substantial number of participants for statistical stability.

The current study manages each of these issues, using data from a 40-year longitudinal study. A combination of exact and propensity score matching was used to reduce the imbalance between grade skipping and nonskipping participants on several important baseline measures. In addition, mathematical talent is a well-known indicator of subsequent STEM accomplishments (Benbow, 1992; Lubinski & Benbow, 2006; National Science Board, 2010a; Super & Bachrach, 1957; Wai, Lubinski, & Benbow, 2009), so, accordingly, we use a large sample of mathematically precocious adolescents.

Before proceeding to the particulars of this study, some considerations from developmental theory are in order because they suggest that we might anticipate sex differences when real-world criteria are employed as outcomes specifically indicative of STEM accomplishments. It has been observed that while mathematically precocious male and female students earn advanced educational credentials at commensurate rates (Lubinski & Benbow, 2006), they tend to do so in different disciplines. Mathematically precocious female students are more likely to earn advanced degrees in the life and social sciences than are male students, whereas the inverse is true for engineering and the inorganic sciences. There is an appreciable amount of evidence to suggest that sex differences in lifestyle preferences and, in particular, for people versus things or organic versus inorganic disciplines plays a major role in structuring these differential outcomes (Ceci & Williams, 2007; Ferriman, Lubinski, & Benbow, 2009; Geary, 2005; Halpern et al., 2007; Schmidt, 2011; Su, Rounds, & Anderson, 2009). Therefore, while we do not anticipate sex differences in earning advanced educational credentials among participants, we do anticipate sex differences on some of the criteria utilized for STEM accomplishments: STEM graduate degrees, STEM patents, and STEM publications.

The Present Study

Three key hypotheses drawn from the time-saving theory were examined. Using data from an observational study of mathematically precocious adolescents tracked longitudinally over 3 decades, the study aims to determine whether those who were grade-based accelerated, or grade skipped, after identification at or before age 13 years (a) were more likely to pursue and earn advanced educational degrees and accomplishments in STEM fields, (b) reached these outcomes earlier than their nonaccelerated, intellectual peers, and (c) were more productive than nonaccelerates when assessed at midcareer. We focus on grade skipping postidentification to allow for matching on covariates measured at initial identification. Any grade skipping prior to identification is treated as a covariate to be matched, as described below.

Two additional design features were used to improve inferences about the long-term effects of grade skipping on STEM outcomes. First, we restricted the sample to those identified as mathematically precocious (top 1%) in early adolescence, which has considerable advantages when investigating the development of career productivity of those in STEM fields. A consistent relationship between mathematical or quantitative abilities and interest and accomplishments in STEM domains has been demonstrated in nationally representative samples (Flanagan et al., 1962; Lubinski, 2010; Wai et al., 2009) and in samples of mathematically precocious individuals (Kell, Lubinski, & Benbow, in press; Lubinski & Benbow, 2006; Lubinski, Webb, Morelock, & Benbow, 2001; Park, Lubinski, & Benbow, 2007, 2008; Wai, Lubinski, & Benbow, 2005). Many of the regional and national indicators of STEM activity (National Science Board, 2010b), such as the number of STEM graduate degrees, peer-reviewed publications, and patents, require large samples for stable results given the low base rate of these indicators in the population.[2]

Second, we used matching to create balanced comparison groups prior to statistical analyses. Several promising new matching methods have been proposed recently (e.g., Diamond & Sekhon, in press; Iacus, King, & Porro, 2012; Ho, Imai, King, & Stuart, 2007; Imai, King, & Stuart, 2008; Sekhon, 2007, 2009), but we combine two popular, longstanding methods, exact and propensity score matching, to create matched samples.

The matching procedure significantly reduces the imbalance between the grade skippers and the resulting control group, allowing estimation of the average effect of the treatment on the treated, or the average effect of grade skipping among those who grade skipped in the following ways. First, we used logistic regression to

[2] Additional descriptive statistics are available from the authors.

adjust further for covariates and ultimately to estimate the effect on the likelihood of earning five indicators of STEM educational or occupational achievement, in the form of incidence or “risk” ratios. Second, we use methods from survival analysis (Harrell, 2001; Kaplan & Meier, 1958; Singer & Willett, 2003) to determine whether grade-skippers reach four educational and occupational accomplishments in young adulthood earlier than their matched and retained intellectual peers. Finally, using the subset of participants who authored STEM publications or earned patents, we compare the citation records and productivity indices of the grade skippers and matched controls to test the hypothesis that grade skipping engenders long-term, cumulative effects.

Method

Sample

Participants were drawn from the first three cohorts of the Study of Mathematically Precocious Youth (SMPY), a planned 50-year longitudinal study of intellectual talent (Lubinski & Benbow, 2006). Each cohort was identified during the intervals 1972–1974, 1976–1979, and 1980–1983 and referred to as the 1972 cohort, 1976 cohort, and 1980 cohort, respectively. Participants in every cohort were identified at or before age 13 years by scores on subtests of the College Board Scholastic Assessment Test (SAT), and each cohort had different but overlapping selection criteria. In the present study, we only included participants from each cohort who scored at or above 390 on the math subtest of the SAT (the SAT Math), which is approximately the lower bound of the top 1% of scores for that age group. Although there is substantial overlap in the entry criteria and the variables measured at the initial assessment for all cohorts, different subsets of background variables were assessed at the initial identification of each cohort.

The 1972 cohort includes 2,188 participants (96% Caucasian, 2% Asian, 2% other) who earned a score of at least 390 on the math subtest of the SAT (the SAT Math) or a 370 on the verbal subtest (the SAT Verbal) by age 13 years. Cut scores denoted the top 1% of this age group, and almost all participants scored beyond these cutoffs. This cohort was drawn primarily from the state of Maryland, with most from the Baltimore–Washington area.

The 1976 cohort includes 778 participants (89% Caucasian, 6% Asian, 5% other) scoring at least 500 on the SAT Math or 430 on the SAT Verbal before age 13 years, the lower bounds of scores of the top 0.5% of this age group. This cohort was drawn primarily from the mid-Atlantic states.

The 1980 cohort includes 501 participants (65% Caucasian, 17% Asian, 1% African American, 1% other, 16% did not disclose) scoring at least 700 on the SAT Math subtest or 630 on the SAT Verbal subtest at or before age 13 years, the lower bounds of scores of the top 0.01% of this age group. This cohort was drawn from talent searches throughout the United States.

After initial identification at or before age 13 years, participants were followed up at ages 18, 23, and 33 years through phone, mail, and Internet surveys (Benbow, Lubinski, Shea, & Eftekhari-Sanjani, 2000; Lubinski, Benbow, Webb, & Bleske-Rechek, 2006). In addition, all participants were followed up with searches of public Internet databases such as the ProQuest Dissertations and Theses database (http://proquest.umi.com), Google Scholar (http://www.google.com/scholar), and Google Patents (http://www.google.com/patents).

Baseline Measures

At the time of identification, participants completed questionnaires about their academic preferences, perceived ability, number of siblings, and their parents’ education and occupations, and these measures were used in the matching process. Several identical items were presented to participants in every cohort, and many typical items are listed in Appendix Table A1. Because every baseline measurement was assessed at the initial participant identification and the grade skipping treatment of interest always took place after identification, all baseline measurements in this study are “pretreatment.”

1972 cohort. Appendix Table B1 lists 14 variables collected at the initial identification of the 1972 cohort. Most participants in this cohort were identified by scores on the Math subtest of the SAT and are missing scores on the Verbal subtest; therefore, only SAT Math scores were used in this study for this cohort.

1976 cohort. Appendix Table B2 lists 21 variables collected at the initial identification of the 1976 cohort. Several variables are identical to those collected in the 1972 cohort with the exceptions described here.

1980 cohort. Appendix Table B3 lists 20 variables collected at the initial identification of the 1980 cohort. Several variables are identical to or similar to those collected in the 1972 and 1974 cohorts.

Missing Data

Some items were introduced after the beginning of the initial assessment procedure, resulting in missing values on these variables for some participants. This problem is mostly confined to the subject matter preference variables in 1972 cohort, and very few observations are missing in the two later cohorts. Missing values were multiply imputed using the Amelia II package in R (Honaker, King, & Blackwell, 2007; Horton & Kleinman, 2007; King, Honaker, Joseph, & Scheve, 2001; Rubin, 2004). Parameter estimates, such as regression coefficients or incidence ratios, are estimated in each dataset and then averaged across datasets to derive point estimates for each parameter using the Zelig package in R (Imai, King, & Lau, 2007, 2009).

In the current study, variables with missing values were used to estimate individual propensity scores, and these scores were in turn used to find well-matching control observations. Consequently, the multiple imputation procedure results in 10 different (but highly overlapping) matched control groups, one in each imputed dataset (Crowe, Lipkovich, & Wang, 2010; Qu & Lipkovich, 2009).[3] All reported statistical summaries in the current analysis combine information from the 10 imputed datasets for each cohort. Means of the baseline measures in the control and grade skipping groups

[3] The imputed control groups were highly similar in their propensity score distributions. To quantify this, we computed the percentage of overlap of the propensity score distributions for every pair of control groups within a cohort using the procedure suggested by Tilton (1937) and noted the amount of overlap between the two least similar distributions. For the 1972 cohort, the two most dissimilar control groups had approximately 98% overlap. These values in the 1976 and 1980 cohorts were 98% and 97%, respectively.

for each cohort prior to missing data imputation are presented in Appendix Tables D1–D3. The numbers of missing observations on each measure are presented in Appendix Tables D4–D6.

Grade Skipping

At ages 18 and 23 years, participants responded to items in follow-up questionnaires concerning the different types of educational acceleration they experienced since the initial assessment. Based on these responses, it was possible to determine the number of grades skipped by each participant. Most participants in all three cohorts did not skip any grades during this period, and those who did skip tended to skip only one full grade. However, some participants did skip more than one grade. Rather than remove those participants, the number of grades skipped after assessment was coded as a dichotomous variable (0 reflecting no grades skipped and 1 for one or more grades skipped). Therefore, the resulting analysis tested directional hypotheses, comparing all grade skippers to the matched nonskippers, rather than estimating the effect of skipping exactly one grade. The total number of grades skipped prior to identification was recorded and used in the match procedure described below.

Additionally, the availability of other types of accelerative opportunities, apart from grade skipping, increased with each subsequent cohort. This was due both to the availability of additional methods of acceleration over time and to the increased level of ability across cohorts. Because most of these opportunities are experienced after identification, they are not included as covariates in the matching procedure.

Matching Procedure

We used a combination of exact and propensity score matching to improve balance between the grade skippers and the comparison group. Propensity score matching (Rosenbaum & Rubin, 1983) counters the multidimensionality of the covariate space by reducing any individual unit's values on observed covariates to a single value between 0 and 1, also known as the propensity score. This score is often interpreted as the probability for receiving treatment, but it can also just stand as a useful summary of an individual's observed values on relevant covariates. We estimate individual propensity scores using logistic regression, predicting grade skipping from the available covariates in each of the three cohorts under analysis in this study. Treated units are then matched with control units based on the propensity score, most commonly using nearest-neighbor matching. This repeats until all treatment units are matched with one control unit, and the process can be repeated, matching additional control units to each treated unit to increase efficiency.

Exact matching matches those treated (in this case, the grade skippers) and control units with exactly the same covariate values. This is highly effective when only one or two covariates are of interest, but exact matching quickly becomes infeasible when the number of covariates grows larger with a finite sample size.

To combine exact and propensity score matching, we first identified two covariates considered to be critical to the outcomes of interest in the following analyses: sex and the number of grades previously skipped before identification by the study. These two covariates are used in the exact matching step of the procedure. The remaining observed covariates varied slightly among cohorts, but each included at least one measure of cognitive ability (SAT Math or SAT Verbal subtest scores at or before age 13) and measures of interest, academic class standing, and perceived importance of various academic subjects, as well as indicators of parental educational attainment and occupational status, number of siblings, and birth order. These covariates are used to estimate each participant's propensity score. After matching participants exactly on sex and number of previous grades skipped, matches were further improved by matching on the nearest neighbor as defined by the propensity score.

Matching was implemented using the MatchIt package in R (Ho, Imai, King, & Stuart, 2011). Propensity scores were estimated with a logistic regression model, using the baseline covariates from each cohort as predictors. Each grade skipper was matched with the control participant with the nearest propensity score (nearest-neighbor matching) who also exactly matched on sex and the number of previous grades skipped. In the 1972 and 1976 cohorts, which had larger sample sizes, we matched in a 2:1 (control to grade skipper) ratio, and used 1:1 matching in the 1980 cohort.

The usual goal of matching is to improve balance and overlap in the covariate distributions of comparison groups in a quasi-experimental study. Perfectly balanced comparison groups will have the same distribution and covariance structure across all covariates, and the severity of imbalance is the degree of departure from this ideal. Currently, no single measure of balance or imbalance is sufficient for assessing the quality of a matching procedure. Following the recommendations of Imai et al. (2008) and Ho et al. (2007), we assessed balance and adjusted our propensity score model specification using a combination of visual aids (histograms and kernel density plots or empirical quantile–quantile plots of propensity scores and individual covariates) and standardized mean differences on relevant covariates, interactions, and squared terms at each iteration.

As a starting point for the propensity score model in each cohort, we used a simple additive model (including only main effects of each predictor) and assessed the resulting covariate balance after matching on the resulting propensity score. In each cohort, this simple model substantially improved balance such that all standardized mean differences among baseline covariates were smaller than 0.25 standard deviations. Next, we attempted to improve balance further within each cohort by respecifying the propensity score model (by removing predictors, adding squared terms, or adding interactions between predictors). These additional model modifications generally improved balance slightly (using standardized mean differences and visual aids as a guide) over the initial simple model, but improvements leveled off quickly. Model modifications stopped when balance was no longer improved across additional iterations.

Figure 1 illustrates the propensity score distributions, before and after matching, from the final matching procedures in each cohort. Resulting means and mean differences on each baseline covariate are available in appendices (Appendix Tables B1, B2, and B3; see Footnote 2). Results indicate full overlap and improved balance of the propensity score distributions after matching, and most importantly, observed mean differences on individual baseline covariates were substantially reduced. After matching, most of the standardized mean differences between the grade skippers and matched

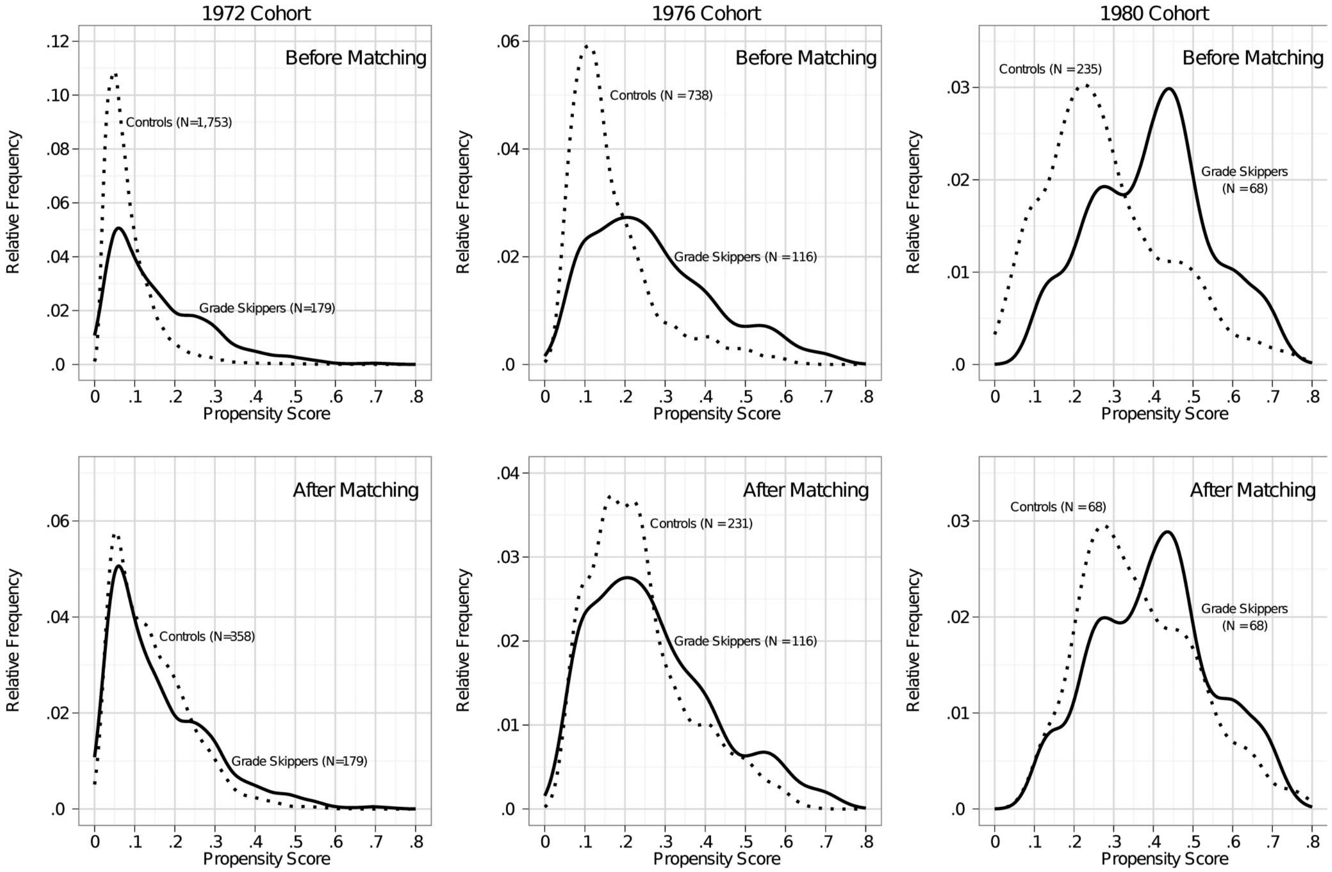

Figure 1. Density plots of propensity score distributions of grade skippers and controls before and after matching. Vertical axes are scaled differently across plots.

controls were smaller than .10, and all were smaller than .25, which has been suggested as the maximum allowed difference to grant the equivalence of randomization to a quasi-experimental or observational design (further adjustments were made using logistic regression, described below; Cochran, 1968; Ho et al., 2007; What Works Clearinghouse, 2009). In addition to propensity score matching, the exact matching constraints ensured that participants were matched exactly on sex and the number of previous grades skipped.

Of the 2,188 participants in the 1972 cohort, 179 (102 male, 77 female) participants were identified as skipping one or more grades, and these were matched with 358 control participants. Matching was most successful in this cohort, in terms of reducing overall mean and distributional imbalances in the observed covariates. The best compromise between sample size and covariate balance was found using a 2:1 (controls to grade skippers) ratio in the matched sample for this cohort. Compared to a 1:1 ratio, using a 2:1 ratio granted a much larger sample size with very little effect on overall balance. For example, using Tilton's (1937) procedure to quantify overlap, the propensity score distributions of the treatment and control groups had approximately 99% overlap using a 1:1 ratio but 97% overlap using a 2:1 ratio. In contrast, using a 3:1 ratio provided an even greater sample size but balance was much worse across most indicators (and overlap dropped to 92%).

From the 778 participants in the 1976 cohort, 116 (97 male, 19 female) participants were identified as skipping one or more grades, and these were matched with 231 control participants. As with the 1972 cohort, a 2:1 ratio was used in the final matched sample. One grade skipper could not be adequately matched with a second control participant, so the initial match for this one participant was duplicated, resulting in 231 matched controls instead of the expected 232.

From the 501 participants in the 1980 cohort, 68 (63 male, 5 female) participants were identified as skipping one or more grades, and these were matched with 68 control participants. To maintain acceptable balance in this cohort, a 1:1 ratio was used in the final matched sample. Still, grade skippers maintained a small average advantage in SAT Math and SAT Verbal scores (approximately .13 and .20 standard deviations, respectively). Nine grade skippers with very high propensity scores in this cohort did not have acceptable matches among the controls. To preserve overlap and balance, these nine grade skippers were dropped from the analysis.

Educational and Occupational Outcomes

Educational degrees. In the current study, only postundergraduate degrees are considered in comparisons (all participants earned undergraduate degrees). Participants in every cohort completed follow-up surveys at age 33, and responses from these surveys were used to determine the educational degrees earned by participants. All participant names were entered into the ProQuest Interdisciplinary Dissertation and Thesis database (http://proquest.umi.com) to determine whether participants completed a dissertation or master's thesis. Any additional information available from participants' professional website or public curriculum vita or resume also was used to determine the educational degrees accumulated by each participant.

Degrees were coded as master's (MA. or MS), PhD, medical degree (MD or equivalent), or law degree (JD). In general, a participant is coded as earning a doctorate if he/she earned a PhD, MD, JD, or a combination of these. Master's and PhD degrees were coded as STEM degrees if they were in the following fields: physical sciences, biological sciences, computer science, engineering, or mathematics. STEM graduate degrees refer to either master's degrees or PhDs from STEM fields.

STEM publications and patents. Every participant name was entered as search terms into Google Scholar to determine whether they were listed as an author on any peer-reviewed publications in scientific journals in STEM fields or listed as an inventor on a granted patent. Matches were confirmed by comparing information from follow-up survey information to the author's or inventor's institutional affiliations. Once a match was confirmed, the total number of publications, patents, and the year of publication of each individual publication or patent were recorded.

Age at accomplishment. By combining birth date information with the month and year of graduation from degree programs or year of publication or granting of a patent, it is possible to estimate a participant's age at the time of reaching each outcome. If both month and year of graduation were available, the age of the participant at graduation was estimated as the number of days between the participant's date of birth and the first day of the month of the graduation year. If only the year of graduation was available, the modal month of observed graduation months was imputed (May). For publications and patents, only the year was available, and age of participant at publication is estimated as the number of days between the date of birth and the middle of the publication year (July 1st). All ages are then converted from days to years by dividing the days by 365.25.

Participant ages at the following four events are used in comparisons: age at graduation from doctoral degree program, age at graduation from STEM graduate degree program, age at publication of first peer-reviewed STEM publication, and age at granting of first patent.

Productivity and citation indices. If participants had at least one citation from a publication or patent, information from the number of publications, the individual citations from each publication, the age of each publication, the total number of citations, and the number of authors on each publication was used to calculate values on a number of common scientific productivity or citation indices. This information was collected using Publish or Perish (POP; Harzing, 2011), software designed to enhance the use of search engines such as Google Scholar. For each participant, four indices were calculated based on their interpretability, robustness, and popularity: total accumulated citations, the *h*-index, the *g*-index, and the age-weighted citation rate.

The *h*-index (Hirsch, 2005), arguably the most popular of all citation and productivity indices, reflects an individual researcher's productivity by combining information about the number of articles they have authored and the number of citations each of those articles has received. According to Hirsch's original definition, "A scientist has index h if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have no more than h citations each" (Hirsch, 2005, p. 1). For example, an *h*-index of 6 means that an individual has published at least 6 articles each with at least 6 citations. This provides a stable metric that is unaffected by "one hit wonder" publications that might heavily skew a raw citation count and favors authors with a steady stream of high-impact articles (Harzing, 2008). As an illustration, Hirsch noted that median *h*-index is 35 among Nobel prize winners and 46 among newly elected members of the National Academy of Sciences in physics and astronomy.

In order to give more weight to heavily cited publications, the *g*-index (Egghe, 2006) is the largest number such that an author's top g articles received together at least g^2 citations. For example, a *g*-index of 15 indicates that an author's top 15 most cited articles together have at least 15^2 or 225 citations, where an *h*-index of 15 indicates that an author's top 15 publications all have at least 15 citations each. Although it is very similar to the *h*-index, relaxing the *h*-index's constraints on distribution citations per article allows the *g*-index to be more sensitive to a skewed distribution of citations across an author's top publications.

The age-weighted citation rate (AWCR; Jin, 2007) reflects the annual rate of citations received by individual's entire body of work, adjusted for the age of each cited publication, calculated by taking the sum of total citations of every publication by an author after dividing the citations from each publication by that publication's age. For example, if an author published 10 articles in the same year, 5 years ago, and each article was cited 20 times, his/her corresponding AWCR would be (20/5)(10), or 40, as this author is cited approximately 40 times per year.

In addition, plotting medians and estimated confidence intervals for the aggregated productivity indices, we assess the difference in the location of the distributions of each index for grade skippers and matched controls using the Wilcoxon Rank Sum test (also known as the Mann-Whitney *U* test; Wilcoxon, 1945). The Wilcoxon test is a nonparametric alternative to a more traditional two-sample *t* test and does not require any distributional assumptions, only the assumption of ordinal scaling.

Results

Educational and Occupational Outcomes

The first step of the analysis was to compare grade skippers and matched controls on the proportions in each group earning advanced educational degrees, STEM publications, and patents. Table 1 lists the percentage of participants in each cohort earning each outcome, as well as percentages pooling across all cohorts. In every comparison, in every cohort, a greater proportion of grade skippers earned doctoral degrees, STEM PhDs, STEM publications, and patents.

Table 1
Percentage Earning Outcome

Cohort and group	*N*	Doctorates	STEM PhDs	STEM publications	Patents
1972 cohort					
Matched controls	358	15.1	3.6	6.4	2.2
Grade skippers	179	27.4	10.1	12.8	4.5
1976 cohort					
Matched controls	231	23.8	14.3	21.2	8.2
Grade skippers	116	31.0	18.1	25.9	9.5
1980 cohort					
Matched controls	68	33.8	17.6	23.5	10.3
Grade skippers	68	45.6	29.4	38.2	17.6
All cohorts					
Matched controls	657	20.1	7.9	13.4	5.2
Grade skippers	363	32.0	16.3	20.9	8.5

Note. Percentages of participants earning outcomes across each cohort and for all cohorts together. The last two columns list the percentage of participants in each category with at least one peer-reviewed publication in a STEM field or patent, respectively. STEM = science, technology, engineering, and mathematics; PhD = doctor of philosophy.

A useful summary for such comparisons is the incidence ratio, also known as the cumulative incidence ratio or risk ratio,[4] interpreted here as the ratio change in average "risk" or the probability of reaching these outcomes due to grade skipping among the grade skippers (Cummings, 2009; Greenland, 1987). Adjusted incidence ratios, which are adjusted for other observed covariates, can be estimated using a logistic regression model, by comparing the average expected values for each participant as the grade skipping variable is changed from 0 to 1. The adjusted incidence ratios (adjusting for all of the background covariates available in each cohort) and 95% confidence intervals for each incidence ratio were estimated and plotted in Figure 2.[5]

In each cohort, the logistic regression model included the dichotomous grade skipping variable and every available baseline covariate as main effects. Using a combination of matching (as a preprocessing step) with a parametric model provides estimates that are "doubly robust" (Ho et al., 2007). With the exception of exact matching, no matching algorithm provides perfect balance, and in the present study, small covariate imbalances remain after the initial matching step. Including these covariates in a parametric model after matching allows additional control without increasing variance substantially because the model is fit to highly balanced data. Additionally, fitting a parametric model to highly balanced data greatly reduces the influence of model specifications such that the effect estimates of the treatment variable are similar across many different model specifications. We use this two-step strategy to estimate incidence ratios within each cohort. We also calculate unadjusted summary incidence ratios across all cohorts and separately for male and female participants. It was not possible to use a second step of parametric modeling for these summary incidence ratios because each cohort had different sets of baseline covariates.

Each outcome has six corresponding estimated incidence ratios, each summarizing the average ratio change in probability in the grade skippers compared to the matched controls. For example, the first three incidence ratios for doctoral degrees are the covariate-adjusted incidence ratios in the 1972, 1976, and 1980 cohorts, respectively. In addition, the next three incidence ratios for doctoral degrees are unadjusted incidence ratios, calculated for all male participants, all female participants, and for all participants across the three cohorts.

An incidence ratio of 1 indicates that there is no difference in the proportions of outcomes across groups. Incidence ratios above 1, or to the right of the dotted vertical line, indicate an increase in the proportion of grade skippers reaching a given outcome. With three exceptions, all point estimates of incidence ratios, in every comparison, is greater than 1. For most individual cohort estimates, 95% confidence intervals around these estimates include 1, indicating that many of the estimates are not statistically significant at the traditional $\alpha = .05$ level. In addition, we summarize the effects across all cohorts, by pooling unadjusted incidence ratios across cohorts for male participants, female participants, and all participants.

Limiting the pooled comparisons only to male participants or female participants reveals an interesting pattern. Results indicate that male grade skippers incurred a much greater increase in the likelihood of earning these outcomes than the female grade skippers, particularly in the comparisons of STEM graduate degrees and STEM PhDs, where female grade skippers were actually less likely than female controls to earn these outcomes. On the other hand, female grade skippers were more likely than their matched controls to earn other doctorates in general.

Table 2 shows the patterns of percentages of different doctoral degrees across male and female participants, grade skippers, and matched controls. The first three subtables show the patterns for each individual cohorts, and the bottom subtable shows the pooled percentages across all three cohorts. The combined percentages show that male grade skippers were much more likely than male controls to pursue STEM graduate degrees and, to a smaller extent, law degrees. Female grade skippers were slightly more likely than female controls to pursue law degrees and medical degrees. After breaking down each subgroup by sex and type of degree, sample sizes in each comparison and the magnitudes of most of the differences are small, but the goal of these comparisons is not to investigate interactions between sex and grade skipping. Rather, these comparisons help explain the seemingly negative effect of grade skipping on female participants based on the incidence ratios in Figure 2. While female participants were less likely to pursue STEM PhDs than male participants, female participants tended to pursue medical degrees at a comparable level and law degrees to

[4] Risk ratios are frequently used in epidemiological contexts to express the change in risk of disease, death, or some other undesirable outcome after exposure or treatment. However, in the current context, an increase in risk is desirable, as the outcomes of interest are accomplishments and generally positive. This led Wai et al. (2010) to use the term "gain ratio" in place of risk ratio when describing the increase in risk of a favorable outcome. We use the neutral terminology incidence ratio. Incidence ratio, cumulative incidence ratio, risk ratio, and gain ratio all have the same interpretation.

[5] For those effect estimates with confidence intervals not containing 1, we performed sensitivity analyses to estimate the robustness of these effects to an unobserved covariate using the rbounds (Keele, 2010) package for R and the methodology proposed by Rosenbaum (2002). For the observed statistically significant effects here, an unobserved covariate could increase the grade skippers' odds of assignment to the treatment group by 1.11–1.34 without changing our inferences. In other words, an unobserved predictor in the logistic regression model with a coefficient greater than .10–.31 (depending on the size of the treatment effect) would be sufficient to explain the observed effects.

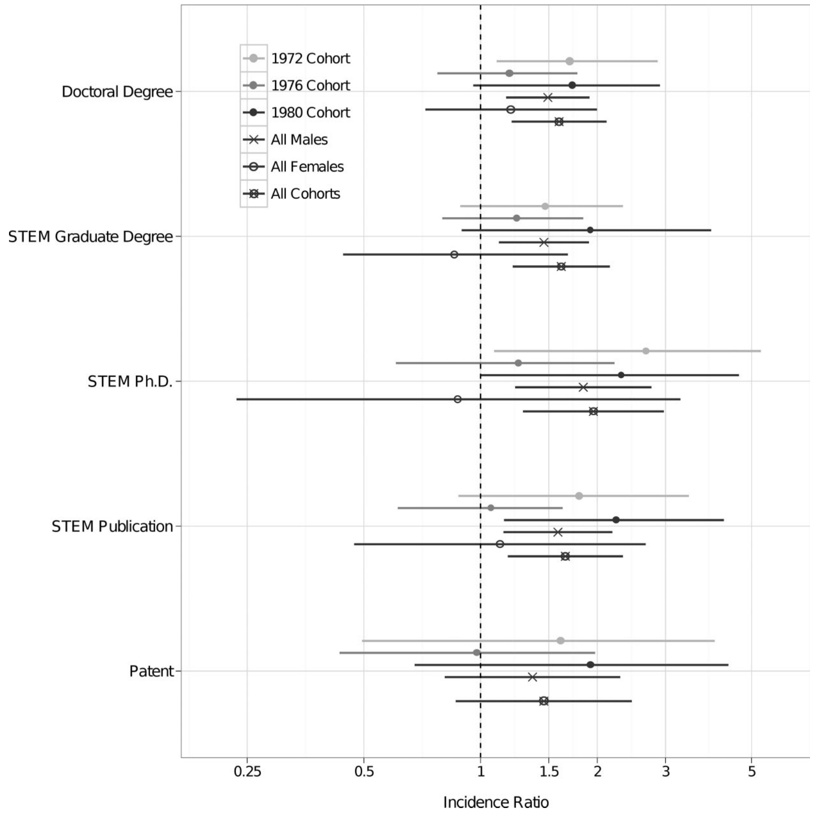

Figure 2. Estimated effect sizes, as incidence ratios, of grade skipping on five outcomes across three cohorts. Points indicate the point estimate of each incidence ratio, and horizontal lines indicate 95% confidence intervals. The vertical line at the incidence ratio of 1 indicates the point of no effect. Incidence ratios for the 1972, 1976, and 1980 cohorts are adjusted for observed covariates. Pooled incidence ratios of all cohorts, all male participants, and all female participants are calculated directly by pooling across each cohort. Note that the incidence ratio for female participants in the patents comparison was not estimated due to the lack of female participants with patents. STEM = science, technology, engineering, and mathematics; PhD = doctor of philosophy.

a greater extent than the male participants did, and these differences were exaggerated among the grade skippers. This tendency, of course, has an impact on STEM publications and patents.

Age at Accomplishment

The next phase of the analysis compared grade skippers and matched controls on the age of occurrence of graduating from a doctoral degree program, graduating from a STEM PhD program, publishing the first STEM publication, and earning the patent. The time-saving theory predicts that grade skippers should reach all outcomes earlier than their matched controls.

Table 3 lists median ages of reaching each outcome among those who did in each cohort and separately for all cohorts pooled together. Median ages are used because the distributions of ages for all outcomes were positively skewed. In the majority of individual comparisons, grade skippers reach the outcomes earlier, and in the pooled comparisons, grade skippers have a median age advantages ranging between .8 (patents) and 1.6 (STEM PhD graduation) years.

Of particular interest is the varying age advantage in authoring the first STEM publication across cohorts. In the 1972 cohort, grade skippers had their first publication at a median age of 25.2, compared to 28 in the matched controls, an advantage of almost 3 years. This advantage shrank to 1.7 years in the 1976 cohort and to .3 years in the 1980 cohort. While the median age of first STEM publication was approximately 25 for grade skippers in all cohorts, the median age for their matched controls steadily decreased across cohorts, and this trend resulted in a shrinking observed advantage.

To illustrate how these differences unfold over time, inverted Kaplan-Meier estimates of survivor functions (Kaplan & Meier, 1958; Singer & Willett, 2003) are shown in Figure 3 (pooled across the three cohorts) and in Appendix Figure C1 (for each individual cohort). Each panel shows the cumulative proportions, in each cohort, of grade skippers and matched controls reaching each outcome as they progress from age 20 years to age 45 years.[6] Median ages within each subgroup (as listed in Table 2) are denoted as vertical lines extending downward from each survivor function. To illustrate the variability in these medians, 95% con-

[6] To maintain consistency across figures, similar horizontal axes are used. However, the median age of the 1980 cohort participants is currently 42 years.

fidence intervals are constructed around each group median using the percentile bootstrap.[7] These intervals are drawn as horizontal line segments passing through the group medians.

Grade skippers tended to reach each outcome earlier, and the median ages of reaching outcomes also tended to decrease across cohorts, with the 1980 cohort reaching many of the outcomes earliest in their lives compared to the other two cohorts. Distributions of doctoral degree graduation and STEM PhD graduation tended to have the smallest variance, with most participants finishing in their mid- to late 20s and very few graduating after age 35 years. The ages of STEM publications and patents showed greater variability, with some participants authoring their first publications while still in their teenage years, but some authoring their first in their late 30s. Patents showed similar variation, shifted even later in life, perhaps reflecting the additional time required to develop a patentable idea.

As with the incidence ratio comparisons, cohorts are pooled in Figure 3 to summarize the findings across cohorts. Similarly, pooled comparisons of ages show consistent age advantages of about 1 year to 1.5 years for doctoral degrees, STEM PhDs, and STEM publications but not patents.

Adult Productivity at Midcareer

The time-saving theory predicts that the time saved from grade skipping, demonstrated in the previous step of the analysis, allows for greater productivity in the long run. Figure 4 shows the relationship between the age of first STEM publication and the total number of citations accrued by participants in all three cohorts. For consistency, horizontal axes are constant across cohorts, but the cohorts differ in their current ages. Total citation counts reflect the total citations received by participants at the time of the most recent measurement in early 2011, when the median ages of the cohorts were 50, 46, and 42. Citations counts followed an approximately log-normal distribution, with many participants having citation counts in the hundreds and a few in the thousands.

Table 2
Percentage Earning Outcome

Cohort and group	*N*	MD	JD	STEM PhD
1972 cohort				
Men	306	8.3	6.6	8.9
Grade skippers	102	6.9	8.8	17.6
Matched controls	204	9.0	5.5	4.6
Women	231	7.8	6.3	2.4
Grade skippers	77	7.8	9.1	0.0
Matched controls	154	7.8	5.0	3.5
1976 cohort				
Men	243	4.9	4.0	19.2
Grade skippers	81	7.4	3.7	21.0
Matched controls	162	3.7	4.2	18.4
Women	104	8.6	8.9	6.5
Grade skippers	35	5.7	8.6	11.4
Matched controls	69	10.0	9.0	4.0
1980 cohort				
Men	126	7.4	4.2	24.0
Grade skippers	63	7.9	4.8	31.7
Matched controls	63	7.0	3.6	16.2
Women	10	21.6	4.8	6.5
Grade skippers	5	40.0	0.0	0.0
Matched controls	5	3.2	9.7	12.9
All cohorts				
Men	675	6.9	5.2	15.4
Grade skippers	246	7.3	6.1	22.4
Matched controls	429	6.7	4.7	11.5
Women	345	8.4	7.1	3.7
Grade skippers	117	8.5	8.5	3.4
Matched controls	228	8.4	6.3	3.9

Note. Percentages of male and female participants earning different doctoral degrees across grade skippers and matched controls. Percentages for the matched controls are averaged over all imputed datasets and do not necessarily represent the percentages in any single imputed dataset. STEM = science, technology, engineering, and mathematics; PhD = doctor of philosophy; MD = doctor of medicine; JD = doctor of jurisprudence.

Table 3
Median Age of Reaching Outcome

Cohort and group	*N*	Doctoral graduation	STEM PhD graduation	First STEM publication	First patent
1972 cohort					
Matched controls	358	26.4	30.1	28.0	37.8
Grade skippers	179	26.2	26.7	25.2	33.7
1976 cohort					
Matched controls	231	27.3	27.8	27.2	35.0
Grade skippers	116	26.9	28.0	25.5	37.2
1980 cohort					
Matched cohort	68	27.1	27.0	26.1	29.8
Grade skippers	68	25.4	26.3	25.8	32.1
All cohorts					
Matched controls	657	27.1	27.8	27.1	35.4
Grade skippers	363	26.3	26.2	25.6	34.6

Note. Median ages (in years) of reaching STEM outcomes, within and across cohorts together. STEM = science, technology, engineering, and mathematics; PhD = doctor of philosophy.

To depict trends within the clouds of points, a nonparametric locally weighted regression (*loess*; Cleveland, 1993; Cleveland & Devlin, 1988) line with a wide bandwidth was fit in each plot, shown in bold. Rather than use all the data and a least-squares estimate of the slope of a single line through it, a loess fit steps across the range of the data, finding the best fit for each portion of the data. To show the stability of these trends, each loess fit was complemented with 100 bootstrap replications, shown by the light grey lines. Each replication fit is created by sampling, with replacement, *n* observations from the original data with sample size *n*, and then fitting the line to that replicated data set. These replicated fits illustrate the robustness of the original fits (in bold) to individual observations.

Plots in Figure 4 show the relationships between the age of a participant at the time of his or her first peer-reviewed STEM publication (*x*-axis) and his or her total citation count at midcareer (scaled logarithmically along the *y*-axis). The general negative trends in all three cohorts indicate that those with earlier first publications tended to have more citations in the long run. The most highly cited participants tended to be those who started

[7] For each subgroup median, confidence intervals were constructed by sampling with replacement from the observed distribution of subgroup ages. For a subgroup with *n* participants reaching an outcome, *n* observations are randomly sampled, with replacement, from the observed distribution of that subgroup's *n* ages, and the median of this age is calculated and recorded. This process is repeated 1,000 times, resulting in 1,000 medians. The 95% confidence intervals are calculated using the values of the 2.5th and 97.5th percentiles of these 1,000 medians.

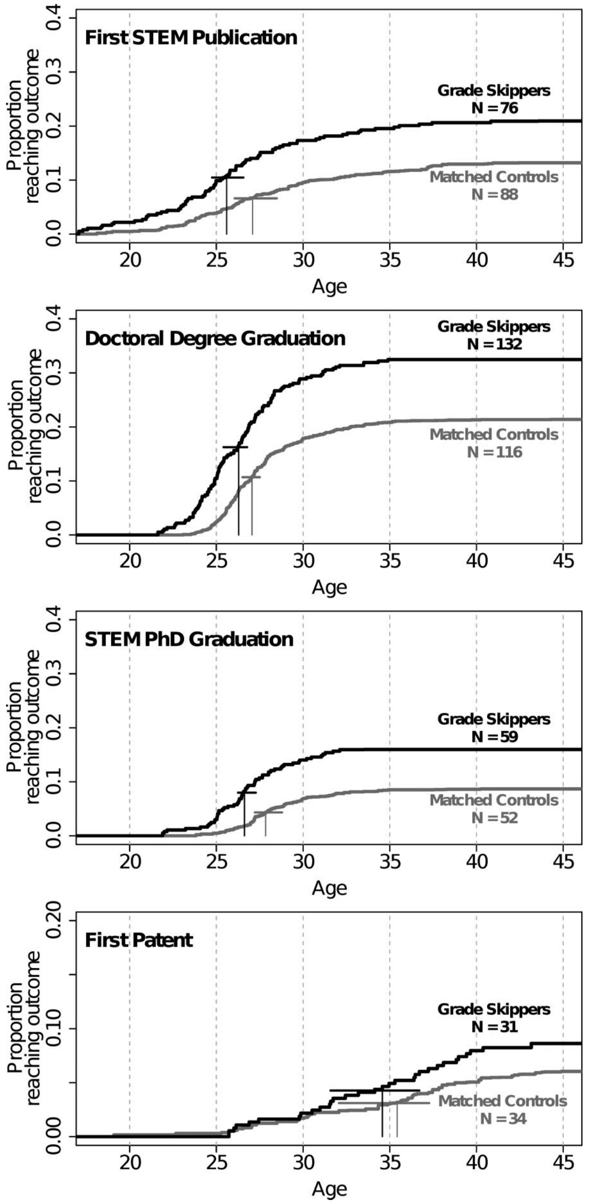

Figure 3. Inverted Kaplan-Meier estimates of survivor functions for four outcomes, pooling all three cohorts together. Vertical line segments indicate the median age of event occurrence for all reaching the event in each group. Horizontal line segments indicate bootstrapped 95% confidence intervals for the medians. STEM = science, technology, engineering, and mathematics; PhD = doctor of philosophy.

publishing in their early 20s. Similar trends are seen if age at STEM PhD graduation is used on the *x*-axis in place of age of first publication.

As shown in Figure 3, the grade skipping participants tended to earn STEM PhDs and author STEM publications earlier than their matched controls. Figure 4 shows that reaching these outcomes at an earlier age was associated to increased productivity, in the form of citations, over the course of participants' careers. The next step is to determine whether the grade skippers were indeed more productive than their matched controls at midcareer, based on similar indices.

Figure 5 plots the differences in median values of four citation and productivity indices for grade skippers and matched cohorts in the 1972, 1976, and 1980 cohorts, respectively, with the left column displaying results from male and female participants and

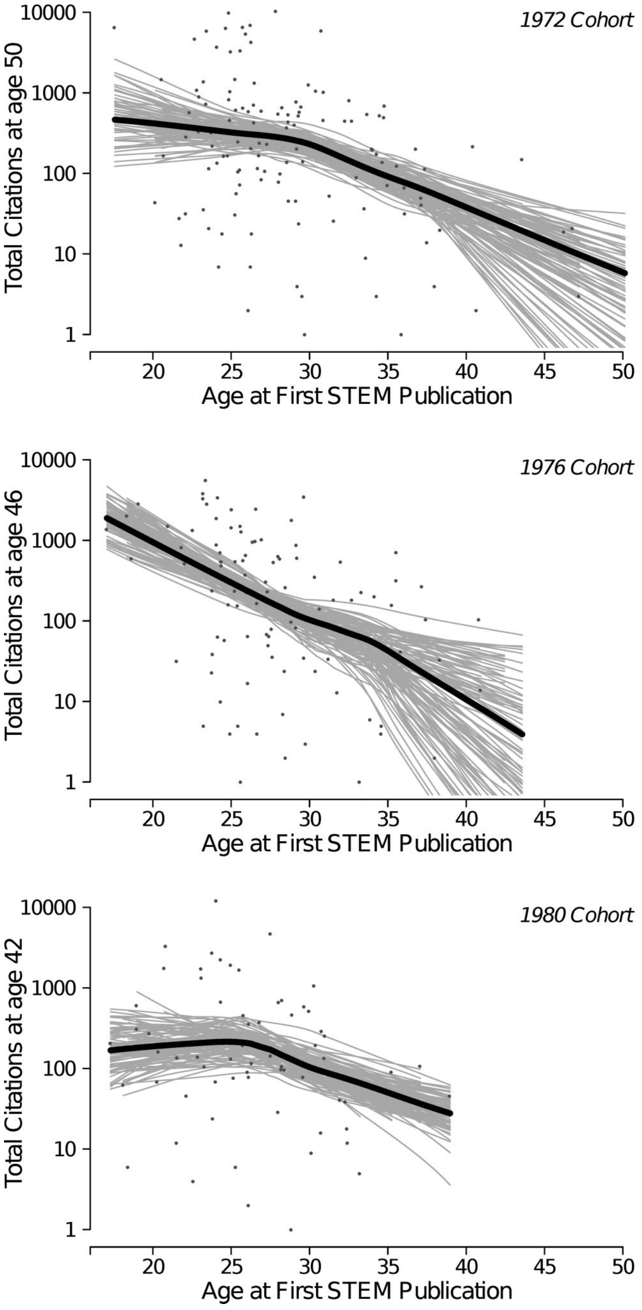

Figure 4. Scatterplots of age at first peer-reviewed STEM publication and total citations (scaled logarithmically). Citation data were collected in 2011 when the 1972, 1976, and 1980 cohorts were 50, 46, and 42 years old, respectively. Black trend lines are fitted using a locally weighted regression (loess), and light grey lines are 100 bootstrap replications of the loess fit. STEM = science, technology, engineering, and mathematics.

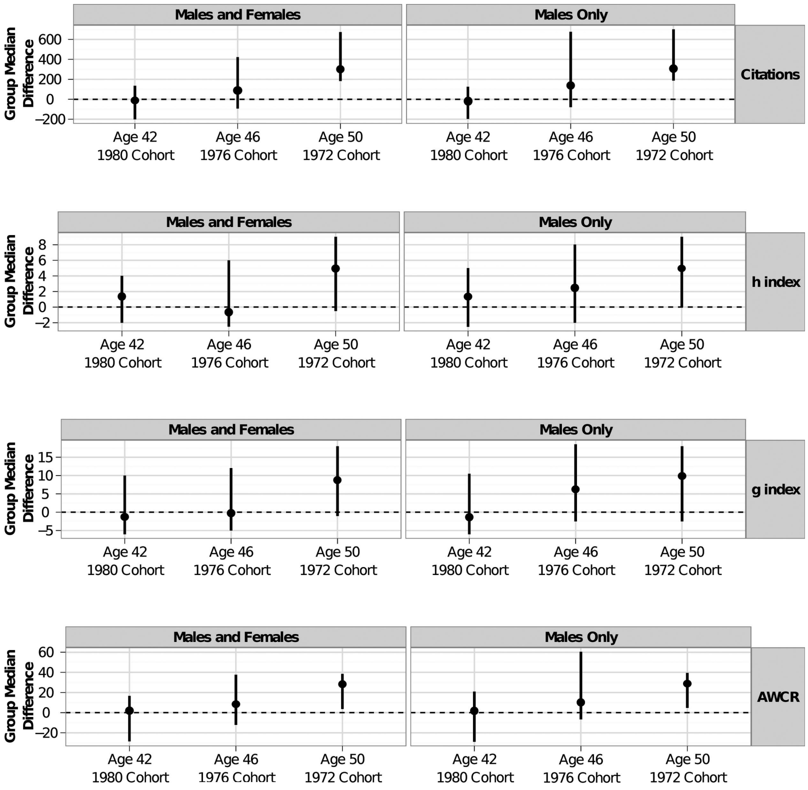

Figure 5. Median differences in productivity/citation indices, comparing grade skippers and matched controls in each cohort. Total citations refers to a participant's total number of citations accumulated from their own peer-reviewed publications and patents. The h-index and g-index are productivity indices based on a combination of a participant's published articles or patents and their respective patterns of citations. A participant with a higher h-index or g-index has authored highly cited articles or earned more highly cited patents than participants with lower values on these indices. The annual citation rate is an estimate of a participant's annual rate of citations, based on the age-weighted citation rate (AWCR). The dashed horizontal line in each plot indicates the point of no group difference. Ages, on the *x*-axis, refer to the median age of the respective cohort at the time of data collection in 2011. Only those participants with at least one citation are included. Confidence intervals around each median difference are estimated using a percentile bootstrap.

the right column restricting the comparison to male participants. Indices include age-weighted citation rate (ACWR) or estimated annual citation rate, the total number of accumulated citations, the *h*-index, and the *g*-index. Only participants with at least one citation can have valid measures on these indices, and many participants excluded from these comparisons had at least one publication but have never been cited.

Unlike the previous steps of this analysis, based on data that are unlikely to change as time passes, the citation and productivity indices are much like snapshots of a process that is continuing to unfold. Indices from the 1972, 1976, and 1980 cohorts were taken when participants were at the median ages of 50, 46, and 42 years, respectively, and these individuals were actively publishing in their respective fields. Because participants in each cohort were at different points in their careers, each cohort is plotted separately and no pooling was done across cohorts.

Inspection of Figure 5 shows a distinct advantage across all indices at age 50 years for the grade skippers. However, similar comparisons from the 1976 and 1980 cohorts are less clear. In the 1976 cohort, grade skippers and their matched controls are similar on most indices at age 46 years, with the matched controls slightly higher. In the 1980 cohort, taken at age 42 years, the opposite pattern is found, with the advantage returning to the grade skippers on most indices.

The grade skippers in the 1976 comparison contain a disproportionately high number of female STEM authors (20.0%) compared to the 1972 (0%) and 1980 (9.7%) cohorts. Moreover, male and female participants across all cohorts tended to have different patterns of publications and citations, with many female participants publishing earlier in their career and less as their career developed. Male participants tended to publish more consistently throughout their careers. To clarify the current comparisons, the right column of Figure 5 displays only male grade skippers and their matched controls. Restricting these comparisons to male participants reveals a pattern of increasing advantage among grade skippers that increases from age 42 years (1980 cohort) to age 46 years (1976 cohort) to age 50 years (1972 cohort).

Two approaches were used to assess the uncertainty in the median differences. First, 95% confidence intervals around each median difference were estimated using a percentile bootstrap, shown as the bands around each median difference in Figure 5. Second, the Wilcoxon Rank Sum test was used to compare median values on each index between grade skippers to their matched controls. Tests were restricted to pairwise comparisons within cohorts for each index. No adjustments for multiple comparisons were made due to the dependent nature across the different indices. To complement the visual comparisons in Figure 5, the ranges of *p*-values from the Wilcoxon Rank Sum tests of differences are reported. While *p*-values are not measures of effect size, they can be a useful guide for assessing the relative magnitude of the differences shown in the figure.

No differences between grade skippers and matched controls at age 42 years (the 1980 cohort), for either combined or male only comparisons, were different according to traditional standards of statistical significance ($.99 > p > .82$ for all eight comparisons). Differences at age 46 years (the 1976 cohort) were also small and insignificant when both male and female participants were included ($.32 > p > .17$). Restricting the comparisons to only male participants increased the magnitude of these differences ($.09 > p > .05$). The largest differences between grade skippers and matched controls were observed at age 50 years (the 1972 cohort). Due to the low proportion of female participants in the original comparisons in this cohort, the magnitude of these differences from the male and female comparisons ($.05 > p > .01$) did not change much when the comparison was restricted to male participants ($.04 > p > .01$).

Discussion

Results from each phase of this study are supportive of key hypotheses of the time-saving theory (Pressey, 1946b), suggesting that grade-based acceleration, appropriately applied with mathematically precocious adolescents (Benbow & Stanley, 1996; Stanley, 2000), can have lasting effects on the productivity of those pursuing careers in STEM fields. The first phase, summarized by Figure 2, reinforces past findings in the acceleration literature (e.g., Bleske-Rechek, Lubinski, & Benbow, 2004; Colangelo, Assouline, & Gross, 2004; Flesher & Pressey, 1955; Kulik & Kulik, 1984; Pressey, 1967; Swiatek & Benbow, 1991; Wai, Lubinski, Benbow, & Steiger, 2010). As in these previous studies of grade-based acceleration, grade skippers were more likely to pursue advanced degrees and secure important career accomplishments related to success in STEM careers, such as STEM publications and patents. The current study not only replicates these findings, but it also strengthens them by revealing similar patterns of results under the much stricter methodological controls granted by the matching procedure.

Recent calls (National Science Board, 2010a) for increasing the STEM workforce and building STEM expertise have focused on both identification and development of national human capital. As shown in Table 1, both the grade skippers and their matched controls earned highly sought achievements in STEM domains at rates several times higher than base expectations. This underscores identification of a population of exceptional promise for STEM accomplishment.

A key finding from the first phase is that although identification of mathematical talent is critical (Park et al., 2007, 2008), interventions based on this identification can enhance development among those with potential for STEM accomplishments. Based on test scores and their responses on background questionnaires at initial identification, the grade skippers were among the most talented and motivated participants. Matching allowed the identification of similarly talented and motivated participants, and these matched controls represent our best estimate of what the grade skippers would be like had they not grade skipped after identification. As shown in Table 1, the matched controls did not flounder without grade skipping. In fact, they earned all of the same accomplishments at very high rates and are clearly at promise for STEM achievement, too. But it appears that a relatively simple intervention, such as grade skipping, can develop this pool of talent even further.

The study's second phase, which focused on the hypothesis that grade skippers would ultimately reach their first STEM accomplishments at earlier ages, extends the findings concerning the effect on age of accomplishments in past literature. Earlier research from SMPY (Stanley, 1973; Swiatek & Benbow, 1991) revealed that participants who skipped grades or entered college early indeed had a time-saving effect that was observable into their early 20s, and accelerated participants tended to finish undergraduate programs and enter graduate programs at an earlier age. At the time, however, participants were not yet old enough to determine whether this effect would last. Currently, virtually all participants in the first three cohorts of SMPY have entered and completed any attempted graduate degrees and are well into their careers, and the results in Figure 3 and Table 2 support the lasting effects of grade skipping. Grade skippers entered and finished their STEM graduate degrees earlier, and similar effects are found when criteria are broadened to include all doctorates and STEM publications.

The finding that grade skippers indeed reach milestones earlier than matched controls fills an existing gap between the educational acceleration literature (e.g., Flesher & Pressey, 1955; Pressey, 1949, 1967; Stanley, 1973; Swiatek & Benbow, 1991) and work on age and lifetime accomplishment (e.g., Dennis, 1956; Lehman, 1946, 1953; Simonton, 1988; Zuckerman, 1977). Many researchers have found a consistent relationship between the age of first accomplishment and the volume of subsequent achievement, but this literature has been almost exclusively retrospective in nature, starting with a highly accomplished individual and working backward to determine the age of their first major accomplishment (Pressey, 1955). Although these studies often lead to fascinating personal histories (Simonton, 1988), age of accomplishment is always confounded with individual differences in ability, motivation, and opportunity.

Figure 4 illustrates the familiar relationship between age of first accomplishment and career productivity within the SMPY sample, using accumulated STEM publications and citations from those publications as indicators. On its own, it is not particularly powerful, but in combination with the findings from the second phase based on this same sample, which show that grade skipping does indeed decrease the age of first accomplishment, the story becomes clearer. The critical piece of this puzzle, showing that the age of accomplishment mediates the effect on later productivity, is arguably still out of reach with observational data (Bullock, Green, & Ha, 2010; Green, Ha, & Bullock, 2010; Zhao, Lynch, & Chen, 2010), but the aggregate findings from all three phases of this study constitute some of the most compelling evidence of the effects of acceleration on adult productivity to date.

The final phase of the study is the first, to our knowledge, to study longitudinal effects of educational acceleration on subsequent STEM accomplishments as fine-grained as citations and citation indices of STEM researchers. Past research (Park et al., 2007, 2008; Wai et al., 2010) used dichotomous outcomes to code whether individuals earned any STEM outcomes or none at all. These criteria are useful in a variety of contexts, but they cannot distinguish between active researchers and inactive researchers or, more importantly, active researchers and prolific researchers. The time-saving theory predicts that if two individuals follow the same career path in STEM, the accelerated participant will be more productive, *ceteris paribus*. To test this theory, indices like citation counts and the *h*-index are useful in distinguishing between levels of productivity among STEM researchers. Looking at research impact is an example of a general issue: Entering a career earlier, does it lead to greater impact?

Narrowing the scope of the analysis to only male participants, for greater clarity, shows a pattern consistent with this interpretation, as seen in the right column of Figure 5. Restricting the comparisons to male participants is reasonable due to the diversity of the paths of the female participants, with many publishing early but later slowing down or transitioning out of research positions into administration or teaching or into entirely different fields or motherhood (Ceci & Williams, 2011; Ceci, Williams, & Barnett, 2009). Career development of talented women seems to follow a different path than that of their male counterparts in many instances.

The results from this phase, summarized in Figure 5, illustrate a pattern of increasing advantage as the cohorts increase in age, such that the grade skippers from the 1980 cohort have no observed advantage at age 42 years while the grade skippers from the 1972 cohort have a significant advantage at age 50 years. Two potential explanations for the observed differences in effect sizes, aside from chance alone, are (a) cohort differences in accelerative opportunities and (b) cumulative effects from grade skipping.

With respect to cohort effects, grade skipping was one of the few accelerative options for the 1972 cohort; the 1976 and especially the 1980 cohorts had many more accelerative opportunities available. The shrinking effect sizes between grade skippers and matched controls in progressively later cohorts may reflect the increased availability of alternative forms of acceleration, such as advanced placement (AP) courses, college courses in high school, summer programs, and research and writing opportunities (Wai et al., 2010), which moderated the differences between the grade skippers and their matches. For example, in the 1972 cohort, the matched controls often reported no other accelerative opportunities, but the matched controls in the 1980 cohort experienced an average of approximately three other forms of acceleration (and on average, just one less opportunity than the grade skippers). In turn, the growth of alternative forms of acceleration over time may explain the progressively smaller effect of grade skipping on age of first STEM publication as well. The 1972 cohort grade skippers tended to author their first publication 3 years earlier than the controls, while the median age advantage in first publication among grade skippers in the 1980 cohort was only 0.3, or about 4 months. While the age of first publication of grade skippers was relatively constant across cohorts, the age of first publication by matched controls gradually decreased across cohorts. It could be that other accelerative opportunities used by the 1976 and 1980 cohorts were almost as effective in saving time as grade skipping. If the effect of grade skipping on these indices is mediated by its effect on age of first publication, then the observed differences across cohorts in Figure 5 are to be expected.[8]

A second explanation is that the grade skipping has small effects that accumulate over time. Assuming that the indices are relatively good "snapshots" of a similar pool of STEM researchers at ages 42, 46, and 50 years, then the gradual increase in the differences between grade skippers and matched controls is the result of the grade skipping advantage. If researchers publish at a relatively constant rate and citation counts grow at an exponential rate (proportional to the amount of publications), then small differences in the time of the first publication will result in gradually widening differences in citation counts as time passes. An idealized example of the process is illustrated in the Appendix Figure C2 using an exponential function to generate accumulated citations from an individual's publication count. The relationship between publications and citations will vary considerably across disciplines and individuals, but the key point is that for any given individual, a small amount of time saved could potentially translate into a large advantage later.

Limitations

A limitation of this study, and matching in general, is that the matching only matches on observed variables, leaving open the possibility that outcome differences between grade skippers and their matched controls were due to differences on unobserved variables and not grade skipping. For example, sensitivity analyses conducted on the effects in Phase 1 (shown in Figure 2) indicate that a small to moderate effect from an unobserved covariate would be sufficient to explain the higher incidence of these outcomes among grade skippers. One would expect that such unobserved covariates would increase (a) the propensity of grade skipping, (b) the likelihood of achieving the rare outcomes studied here, and (c) the age at which they are achieved. While the sets of baseline covariates used in the current study rule out some of the usual purported causal influences, this study cannot definitely rule out other plausible influences that reasonable investigators could posit; indeed, it is possible that unobserved variables are entirely responsible for the observed effects. Therefore, our findings are best thought of as highly suggestive.

Incorporating assessments of spatial ability (Wai et al., 2009), vocational interests (Su et al., 2009), and changes in life priorities following the completion of formal education (Ferriman et al., 2009) would facilitate not only better matching but also the incorporation of a broader range of outcome measures. Similarly, while the 1976 and 1980 cohort matching included school type (public vs. private) as a covariate in the propensity score, more detailed information about participants' school types may improve comparisons, particularly among the roughly 20% in each cohort who were not enrolled in public schools.

Additionally, matching removed hundreds of observations, trading statistical power and precision for a reduction in bias, and this lead to greater uncertainty around the size of many effect estimates. Particularly for female participants in the latter two cohorts

[8] Our positive findings for grade skipping should not be interpreted as if grade skipping is essential for the optimal development of mathematically precocious youth. Indeed, over time, many interventions have developed such that there are multiple ways to meet the needs of intellectually precocious youth, and some intervention modalities may be functionally equivalent (Wai et al., 2010). Just as educational efficacy of an intervention may be compromised by not taking into account the individuality of the student body (Benbow & Stanley, 1996; Bleske-Rechek et al., 2004; Lubinski, 1996, 2010), innovative educational interventions may replace or be used interchangeably with preexisting procedures. The important thing is to treat all students as individuals, and tailor procedures with a keen awareness of the multidimensionality found in each student's individuality (Achter, Lubinski, & Benbow, 1996).

(on both sample size and outcomes), this investigation was limited in its ability to address meaningfully the interactions between sex and grade skipping.

On the other hand, in the spirit of Tukey (1962), it may be better to have a less precise estimate of the correct quantity than a very precise estimate of the wrong one. The imprecision around the effect estimates and the lack of traditional statistical hypothesis testing methods was countered in this study by replicating the broad findings from similar comparisons across three cohorts, following Lykken's (1968) recommendations on the chief importance of replication, relative to statistical significance testing, for evaluating substantive theories.

Another limitation is breadth of the high-accomplishment outcome criteria utilized in this study. Although there is evidence to suggest that mathematically precocious male and female participants achieve at commensurate rates in educational and occupational settings (Lubinski & Benbow, 2006), there is also evidence to suggest that they do so in contrasting areas. Female participants are more likely to focus their creative and occupational energies on organic or life-centered domains, relative to male participants, whereas male participants in turn are more likely to focus on accomplishments in inorganic disciplines (Su et al., 2009). Because there are multiple ways for mathematically precocious students to manifest exceptional accomplishments, outcome criteria should ideally cover a broad range, both inside and outside of STEM arenas. Reflecting on many frameworks in developmental theory, to adequately capture the lifespan accomplishments of intellectually precocious youth, heterogeneous outcome criteria are needed (Ceci & Williams, 2007; Ferriman et al., 2009; Geary, 2005; Halpern et al., 2007; Schmidt, 2011; Su et al., 2009), otherwise the contributions of one or both sexes are likely to be underappreciated (cf. Lubinski & Benbow, 2006).

Finally, we should point out that grade skipping is but one example of a broader category of what has become known as appropriate developmental placement (Lubinski & Benbow, 2000). That is, placing students in learning environments as a function of their readiness to educationally profit from them. Grade skipping is one of many ways to accomplish this for intellectually talented youth; many others exist (Colangelo, Assouline, & Gross, 2004). This is important because although a strong point of the implemented design is its longitudinal time frame, which covers multiple decades, such protracted intervals always have the shortcoming that advances developed subsequent to Time 1 data collection are not initially assessed. There are many interventions that today would be considered complementary to grade skipping and, in some instances, even preferable (Benbow & Stanley, 1996). Yet, like grade skipping, all of these interventions share a common property: the concept of appropriate developmental placement. This more general concept is what is being evaluated to facilitate subsequent accomplishments among intellectually talented youth at the time they were schooled, other things being equal.

Interventions, like assessments of individual differences attributes as well as outcome criteria in general, are best conceptualized as constructs (Cronbach, 1989); while the positive findings reported here do not dictate the implementation of this specific practice (without considering other interventions now readily available), they do suggest that the more general class of interventions from which grade skipping was drawn, namely, appropriate developmental placement, appears to have merit. This idea may be extended to other interventions evaluated in longitudinal research because the power of multiple-decade longitudinal designs is offset, to some extent, by the datedness of the Time 1 interventions initially assessed. However, by conceptualizing such longitudinal inquiry in the context of *constructive replications* (Lykken, 1968), whereby interventions and criteria are being evaluated as indicators of constructs indicative of more general principles, focus returns to the underlying communality cutting across predictor/outcome relationships (cf. Wai et al., 2009), and support for broader generalizations becomes possible.

Summary and Conclusion

Overall, the findings from this study are supportive both of the theory concerning the time-saving mechanism underlying the effects of grade skipping, as described during the peak of interest in acceleration almost 60 years ago (Paterson, 1957; Pressey, 1946b, 1949; Terman, 1954), and also of the more recent policy recommendations following earlier empirical support of acceleration (Benbow & Stanley, 1996; Colangelo, Assouline, & Gross, 2004; Stanley & Benbow, 1982). Mathematically precocious students who grade skipped were more likely to pursue advanced degrees and secure STEM accomplishments, reached these outcomes earlier, and accrued more citations and highly cited publications in STEM fields than their matched and retained intellectual peers.

References

Achter, J. A., Lubinski, D., & Benbow, C. P. (1996). Multipotentiality among the intellectually gifted: "It was never there and already it's vanishing." *Journal of Counseling Psychology, 43,* 65–76. doi:10.1037/0022-0167.43.1.65

Benbow, C. P. (1992). Academic achievement in mathematics and science between ages 13 and 23: Are there differences among students in the top one percent of mathematical ability? *Journal of Educational Psychology, 84,* 51–61. doi:10.1037/0022-0663.84.1.51

Benbow, C. P., Lubinski, D., Shea, D. L., & Eftekhari-Sanjani, H. (2000). Sex differences in mathematical reasoning ability at age 13: Their status 20 years later. *Psychological Science, 11,* 474–480. doi:10.1111/1467-9280.00291

Benbow, C. P., & Stanley, J. C. (1996). Inequity in equity: How "equity" can lead to inequity for high-potential students. *Psychology, Public Policy, and Law, 2,* 249–292. doi:10.1037/1076-8971.2.2.249

Bleske-Rechek, A., Lubinski, D., & Benbow, C. P. (2004). Meeting the educational needs of special populations: Advanced placement's role in developing exceptional human capital. *Psychological Science, 15,* 217–224. doi:10.1111/j.0956-7976.2004.00655.x

Bullock, J. G., Green, D., & Ha, S. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology, 98,* 550–558. doi:10.1037/a0018933

Ceci, S. J., & Williams, W. M. (Eds.). (2007). *Why aren't more women in science? Top researchers debate the evidence*. Washington, DC: American Psychological Association. doi:10.1037/11546-000

Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *PNAS: Proceedings of the National Academy of Sciences of the United States, 108,* 3157–3162. doi:10.1073/pnas.1014871108

Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin, 135,* 218–261. doi:10.1037/a0014412

Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

Cleveland, W., & Devlin, S. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association, 83,* 596–610.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24,* 295–313. doi:10.2307/2528036

Colangelo, N., Assouline, S. G., & Gross, M. U. M. (2004). *A nation deceived: How schools hold back America's brightest students*. Iowa City: University of Iowa.

Colangelo, N., Assouline, S. G., & Lupkowski-Shoplik, A. E. (2004). Whole-grade acceleration. In N. Colangelo, S. G. Assouline, & M. U. M. Gross (Eds.), *A nation deceived: How schools hold back America's brightest students* (pp. 77–86). Iowa City: University of Iowa Press.

Cronbach, L. J. (1989). Construct validity after thirty years. In R. L. Linn (Eds.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana: University of Illinois Press.

Crowe, B. J., Lipkovich, I. A., & Wang, O. (2010). Comparison of several imputation methods for missing baseline data in propensity scores analysis of binary outcome. *Pharmaceutical Statistics, 9,* 269–279. doi:10.1002/pst.389

Cummings, P. (2009). The relative merits of risk ratios and odds ratios. *Archives of Pediatrics and Adolescent Medicine, 163,* 438–445. doi:10.1001/archpediatrics.2009.31

Dennis, W. (1956). Age and productivity among scientists. *Science, 123,* 724–725. doi:10.1126/science.123.3200.724

Diamond, A., & Sekhon, J. S. (in press). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*.

Domestic Policy Council & Office of Science and Technology Policy. (2006). *American Competitiveness Initiative: Leading the world in innovation* [Technical report]. Washington, DC: Executive Office of the President of the United States.

Duncan, O. D. (1961). A socioeconomic index for all occupations. In J. Reiss, Jr. (Ed.), *Occupations and social status* (pp. 109–138). New York, NY: Free Press of Glencoe.

Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics, 69,* 131–152. doi:10.1007/s11192-006-0144-7

Ferriman, K., Lubinski, D., & Benbow, C. P. (2009). Work preferences, life values, and personal views of top math/science graduate students and the profoundly gifted: Developmental changes and sex differences during young adulthood and parenthood. *Journal of Personality and Social Psychology, 97,* 517–532. doi:10.1037/a0016030

Flanagan, J. C., Dailey, J. T., Shaycoft, M. F., Gorham, W. A., Orr, D. B., & Goldberg, I. (1962). *Design for a study for American youth*. Boston, MA: Houghton Mifflin.

Flesher, M., & Pressey, S. (1955). War-time accelerates ten years after. *Journal of Educational Psychology, 46,* 228–238. doi:10.1037/h0041914

Friedman, T. L. (2005). *The world is flat*. New York, NY: Farrar, Straus, & Giorux.

Geary, D. C. (2005). *The origin of the mind: Evolution of brain, cognition, and general intelligence*. Washington, DC: American Psychological Association.

Geary, D. C. (2010). *Male, female: The evolution of human sex differences* (2nd ed.). Washington, DC: American Psychological Association. doi:10.1037/12072-000

Green, D. P., Ha, S. E., & Bullock, J. G. (2010). Enough already about "black box" experiments: Studying mediation is more difficult than most scholars suppose. *The Annals of the American Academy of Political and Social Science, 628,* 200–208. doi:10.1177/0002716209351526

Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology, 125,* 761–768.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8,* 1–51.

Harrell, F. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York, NY: Springer.

Harzing, A. W. (2008). *Reflections on the h-index*. Retrieved from http://www.harzing.com/pop-hindex.htm

Harzing, A. W. (2011). Publish or perish (Version 3.0.4084) [Computer software]. Retrieved from www.harzing.com/pop.htm

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *PNAS: Proceedings of the National Academy of Sciences of the United States, 102,* 16569–16572. doi:10.1073/pnas.0507655102

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15,* 199–236. doi:10.1093/pan/mpl013

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42,* 1–28.

Hobbs, N. (1951). Community recognition of the gifted. In P. Witty (Ed.), *The gifted child* (pp. 163–183). Boston, MA: Heath.

Honaker, J., King, G., & Blackwell, M. (2007). Amelia II: A program for missing data (Version 1.5-2) [Computer software]. Retrieved from http://gking.harvard.edu/amelia/

Horton, N. J., & Kleinman, K. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *American Statistician, 61,* 79–90. doi:10.1198/000313007X172556

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis, 20,* 1–24.

Imai, K., King, G., & Lau, O. (2007). logit: Logistic regression for dichotomous dependent variables. (Version 3.3.5) [Computer software]. Retrieved from http://gking.harvard.edu/zelig

Imai, K., King, G., & Lau, O. (2009). *Zelig: Everyone's statistical software* [Computer software manual]. Retrieved from http://gking.harvard.edu/zelig/docs/zelig.pdf

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A. Statistics in Society, 171,* 481–502. doi:10.1111/j.1467-985X.2007.00527.x

Jin, B. (2007). The AR-index: Complementing the h-index. *International Society for Scientometrics and Informetrics Newsletter, 3,* 6.

Kaplan, E., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53,* 457–481.

Keele, L. (2010). *An overview of rbounds: An R package for Rosenbaum bounds sensitivity analysis with matched data*. Retrieved from http://www.personal.psu.edu/ljk20/rbounds%20vignette.pdf

Kell, H. J., Lubinski, D., & Benbow, C. P. (in press). Those who rise to the top: Early indicators. *Psychological Science*.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *The American Political Science Review, 95,* 49–69.

Kulik, J. A., & Kulik, C. C. (1984). Effects of accelerated instruction on students. *Review of Educational Research, 54,* 409–425.

Lehman, H. C. (1946). Age of starting to contribute versus total creative output. *Journal of Applied Psychology, 30,* 460–480. doi:10.1037/h0054009

Lehman, H. (1953). *Age and achievement*. Princeton, NJ: Princeton University Press.

Lubinski, D. (1996). Applied individual differences research and its quantitative methods. *Psychology, Public Policy, and Law, 2,* 187–203. doi:10.1037/1076-8971.2.2.187

Lubinski, D. (2010). Neglected aspects and truncated appraisals in vocational counseling: Interpreting the interest-efficacy association from a broader perspective: Comment on Armstrong and Vogel (2009). *Journal of Counseling Psychology, 57,* 226–238. doi:10.1037/a0019163

Lubinski, D., & Benbow, C. P. (2000). States of excellence. *American Psychologist, 55,* 137–150.

Lubinski, D., & Benbow, C. P. (2006). Study of mathematically precocious youth after 35 years: Uncovering antecedents for the development of math–science expertise. *Perspectives on Psychological Science, 1,* 316–345. doi:10.1111/j.1745-6916.2006.00019.x

Lubinski, D., Benbow, C. P., Webb, R. M., & Bleske-Rechek, A. (2006). Tracking exceptional human capital over two decades. *Psychological Science, 17,* 194–199. doi:10.1111/j.1467-9280.2006.01685.x

Lubinski, D., Webb, R. M., Morelock, M. J., & Benbow, C. P. (2001). Top 1 in 10,000: A 10-year follow-up of the profoundly gifted. *Journal of Applied Psychology, 86,* 718–729. doi:10.1037/0021-9010.86.4.718

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159. doi:10.1037/h0026141

National Science Board. (2010a). *Preparing the next generation of STEM innovators: Identifying and developing our nation's human capital* (Technical Report NSB 10–33). Arlington, VA: National Science Foundation.

National Science Board. (2010b). *Science and engineering indicators 2010* (Technical Report NSB 10–01). Arlington, VA: National Science Foundation.

Park, G., Lubinski, D., & Benbow, C. P. (2007). Contrasting intellectual patterns for creativity in the arts and sciences: Tracking intellectually precocious youth over 25 years. *Psychological Science, 18,* 948–952. doi:10.1111/j.1467-9280.2007.02007.x

Park, G., Lubinski, D., & Benbow, C. P. (2008). Ability differences among people who have commensurate degrees matter for scientific creativity. *Psychological Science, 19,* 957–961. doi:10.1111/j.1467-9280.2008.02182.x

Paterson, D. G. (1957). The conservation of human talent. *American Psychologist, 12,* 134–144. doi:10.1037/h0043853

Pressey, S. L. (1946a). Acceleration: Disgrace or challenge? *Science, 104,* 215–219. doi:10.1126/science.104.2697.215

Pressey, S. L. (1946b). Time-saving in professional training. *American Psychologist, 1,* 324–329. doi:10.1037/h0062575

Pressey, S. L. (1949). *Educational acceleration: Appraisals and basic problems*. Columbus: The Ohio State University.

Pressey, S. L. (1955). Concerning the nature and nurture of genius. *Scientific Monthly, 81,* 123–129.

Pressey, S. L. (1967). "Fordling" accelerates ten years after. *Journal of Counseling Psychology, 14,* 73–80. doi:10.1037/h0024234

Qu, Y., & Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine, 28,* 1402–1414. doi:10.1002/sim.3549

Rogers, K. B. (2007). Lessons learned about educating the gifted and talented: A synthesis of the research on educational practice. *Gifted Child Quarterly, 51,* 382–396. doi:10.1177/0016986207306324

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55. doi:10.1093/biomet/70.1.41

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Schmidt, F. L. (2011). A theory of sex differences in technical aptitude and some supporting evidence. *Perspectives on Psychological Science, 6,* 560–573. doi:10.1177/1745691611419670

Seashore, C. E. (1922). The gifted student and research. *Science, 56,* 641–648. doi:10.1126/science.56.1458.641

Sekhon, J. S. (2007). *Alternative balance metrics for bias reduction in matching methods for causal inference*. Retrieved from http://sekhon.berkeley.edu/papers/SekhonBalanceMetrics.pdf

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science, 12,* 487–508. doi:10.1146/annurev.polisci.11.060606.135444

Simonton, D. K. (1988). Age and outstanding achievement: What do we know after a century of research? *Psychological Bulletin, 104,* 251–267. doi:10.1037/0033-2909.104.2.251

Simonton, D. K. (1997). Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review, 104,* 66–89. doi:10.1037/0033-295X.104.1.66

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Stanley, J. C. (1973). Accelerating the educational progress of intellectually gifted youths. *Educational Psychologist, 10,* 133–146. doi:10.1080/00461527309529108

Stanley, J. C. (2000). Helping students learn only what they don't already know. *Psychology, Public Policy, and Law, 6,* 216–222. doi:10.1037/1076-8971.6.1.216

Stanley, J. C., & Benbow, C. P. (1982). Educating mathematically precocious youths: Twelve policy recommendations. *Educational Researcher, 11,* 4–9.

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin, 135,* 859–884. doi:10.1037/a0017364

Super, D. E., & Bachrach, P. B. (1957). *Scientific careers and vocational development theory: A review, a critique and some recommendations*. New York, NY: Columbia University, Teachers College, Bureau of Publications.

Swiatek, M. A., & Benbow, C. P. (1991). Ten-year longitudinal follow-up of ability-matched accelerated and unaccelerated gifted students. *Journal of Educational Psychology, 83,* 528–538. doi:10.1037/0022-0663.83.4.528

Terman, L. M. (1954). The discovery and encouragement of exceptional talent. *American Psychologist, 9,* 221–230. doi:10.1037/h0060516

Tilton, J. W. (1937). The measurement of overlapping. *Journal of Educational Psychology, 28,* 656–662. doi:10.1037/h0053750

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics, 33,* 1–67. doi:10.1214/aoms/1177704711

Wai, J., Lubinski, D., & Benbow, C. P. (2005). Creativity and occupational accomplishments among intellectually precocious youth: An age 13 to age 33 longitudinal study. *Journal of Educational Psychology, 97,* 484–492. doi:10.1037/0022-0663.97.3.484

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology, 101,* 817–835. doi:10.1037/a0016127

Wai, J., Lubinski, D., Benbow, C. P., & Steiger, J. S. (2010). Achievement in science, technology, engineering, and mathematics and its relationship to STEM educational dose: A 25-year longitudinal study. *Journal of Educational Psychology, 102,* 860–871. doi:10.1037/a0019454

What Works Clearinghouse. (2009). *What Works Clearinghouse procedures and standards handbook* (Version 2.0). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1,* 80–83. doi:10.2307/3001968

Zhao, X., Lynch, J., Jr., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research, 37,* 197–206. doi:10.1086/651257

Zuckerman, H. (1977). *Scientific elite: Nobel laureates in the United States*. New York, NY: Free Press.

(*Appendices follow*)

Appendix A

Baseline Item Descriptions

Table A1 presents descriptions of items from questionnaires at the initial identification of participants.

Table A1
Descriptions of Baseline Items

Baseline measure	Description (minimum and maximum)
Parental highest degree	Ordinal scale of highest degree earned (In cohort 1972, 1 = *Less than high school*, 7 = *Doctoral degree*; In other cohorts, 1 = *Less than high school*, 9 = *Postdoctoral experience*)
Parental occupational prestige	Occupational prestige according to Duncan (1961; 1 = *minimum prestige*, 100 = *maximum prestige*)
Birth order	Birth order among siblings (1 = *first born*, 10 = *10th born*)
Number of siblings	Number of siblings (0 = *only child*, 10 = *10 siblings*)
Liking for *X*	"What word best describes your liking for *X* ?" (1 = *strongly unfavorable*, 5 = *strongly favorable*)
Doing well in *X*	"Compared to classmates, how well are you doing in your *X* class?" (1 = *less well than most*, 5 = *better than all*)
Learning in *X*	"How are you learning most of your *X*?" (1 = *with my classmates in school*, 4 = *on my own with little help*)
X importance	"How important will *X* be for a job someday?" (1 = *not at all*, 4 = *very important*)

Note. Typical items from initial assessment questionnaires and minimum and maximum possible responses and meaning. Item responses were used in the matching procedure, along with SAT subtest scores and other demographic variables. *X* is a placeholder for various academic subjects used in the items, such as mathematics, biology, or English.

Appendix B

Means of Baseline Measures

Tables B1, B2, and B3 present means on baseline measures for all control participants, matched control participants, and grade skipping participants.

Table B1
Baseline Means of 1972 Cohort

1972 cohort	All controls	Matched controls	Grade skippers
N	1,753	358	179
SAT Math score	517	559	568
Mother's highest degree	3.3	3.6	3.7
Father's highest degree	4.3	4.5	4.5
Mother's occupational prestige	74	75	74
Father's occupational prestige	77	78	78
Birth order	2.1	2.0	2.0
Number of siblings	2.4	2.2	2.3
Liking for school	3.1	3.1	3.2
Liking for math class	3.4	3.5	3.5
Doing well in math class	2.9	3.0	3.0
Learning math	1.3	1.4	1.4
Math importance	4.4	4.4	4.4
Previous grades skipped	0.1	0.2	0.2
Proportion male	0.62	0.57	0.57

Note. Means and proportions of 14 background variables measured at age 13 years across unmatched controls, propensity score matched controls, and accelerates in the 1972 cohort. Liking for school, liking for math class, doing well in math class, learning math, and math importance refer to items presented to participants at their initial identification.

(*Appendices continue*)

Table B2
Baseline Means of 1976 Cohort

1976 cohort	All controls	Matched controls	Grade skippers
N	507	231	116
SAT Math score	548	570	577
SAT Verbal score	455	471	482
Mother's highest degree	4.5	4.7	4.7
Father's highest degree	5.2	5.4	5.4
Number of siblings	1.8	1.7	1.8
Liking for school	3.9	4.0	3.9
Liking for math class	4.3	4.4	4.4
Liking for biology class	3.5	3.5	3.5
Liking for chemistry class	3.8	3.9	3.9
Liking for physics class	3.6	3.7	3.8
Doing well in math class	1.9	1.8	1.8
Doing well in science class	2.1	2.0	1.9
Learning math	1.3	1.5	1.6
Learning science	1.2	1.2	1.2
Math importance	3.5	3.6	3.6
Biology importance	2.6	2.4	2.4
Chemistry importance	2.7	2.8	2.8
Physics importance	2.8	3.1	3.2
Previous grades skipped	0.1	0.2	0.2
Proportion male	0.7	0.7	0.7
Proportion in public school	0.82	0.83	0.84

Note. Means and proportions of 21 background variables measured at age 13 years across unmatched controls, propensity score matched controls, and accelerates in the 1976 cohort. Liking, doing well, and importance variables refer to items presented to participants at their initial identification.

Table B3
Baseline Means of 1980 Cohort

1980 cohort	All controls	Matched controls	Grade skippers
N	167	68	68
SAT Math score	682	716	721
SAT Verbal score	549	541	560
Mother's highest degree	6.0	5.8	5.8
Father's highest degree	7.0	6.9	6.7
Mother's occupational prestige	70	68	68
Father's occupational prestige	80	79	81
Number of siblings	1.4	1.4	1.4
Liking for school	3.8	3.9	4.0
Liking for math class	4.6	4.9	4.8
Liking for biology class	3.7	3.7	3.8
Liking for chemistry class	4.1	4.3	4.2
Liking for physics class	4.2	4.4	4.4
Liking for English class	3.8	3.7	3.9
Liking for writing	3.6	3.3	3.4
Liking for foreign language class	4.1	4.0	4.0
Liking for social studies	3.6	3.7	3.8
Learning math	1.4	1.5	1.7
Previous grades skipped	0.5	0.3	0.3
Proportion male	0.74	0.93	0.93
Proportion in public school	0.79	0.84	0.85

Note. Means and proportions of 20 background variables measured at age 13 years across unmatched controls, propensity score matched controls, and accelerates in the 1980 cohort. Liking, doing well, and importance variables refer to items presented to participants at their initial identification.

(*Appendices continue*)

Appendix C

Additional Figures

Figure C1 displays each cohort's survivor functions, which were used to create the pooled survivor functions in Figure 3. Figure C2 displays a hypothetical data generating mechanism to explain between cohort differences in grade skipping effects on citation indices.

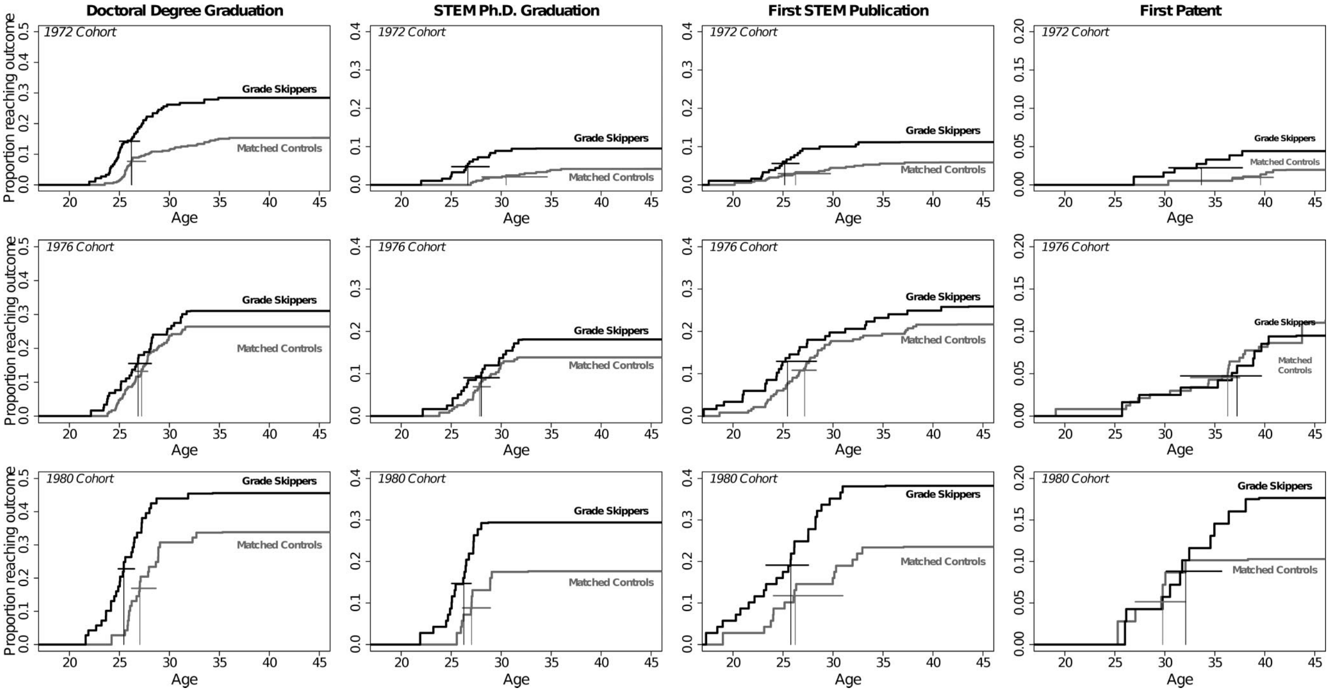

Figure C1. Inverted Kaplan-Meier estimates of survivor functions for four outcomes within each of the three cohorts. Vertical line segments indicate the median age (in years) of event occurrence for all reaching the event in each group. Horizontal line segments indicate bootstrapped 95% confidence intervals for the medians. STEM = science, technology, engineering, and mathematics; PhD = doctor of philosophy.

(*Appendices continue*)

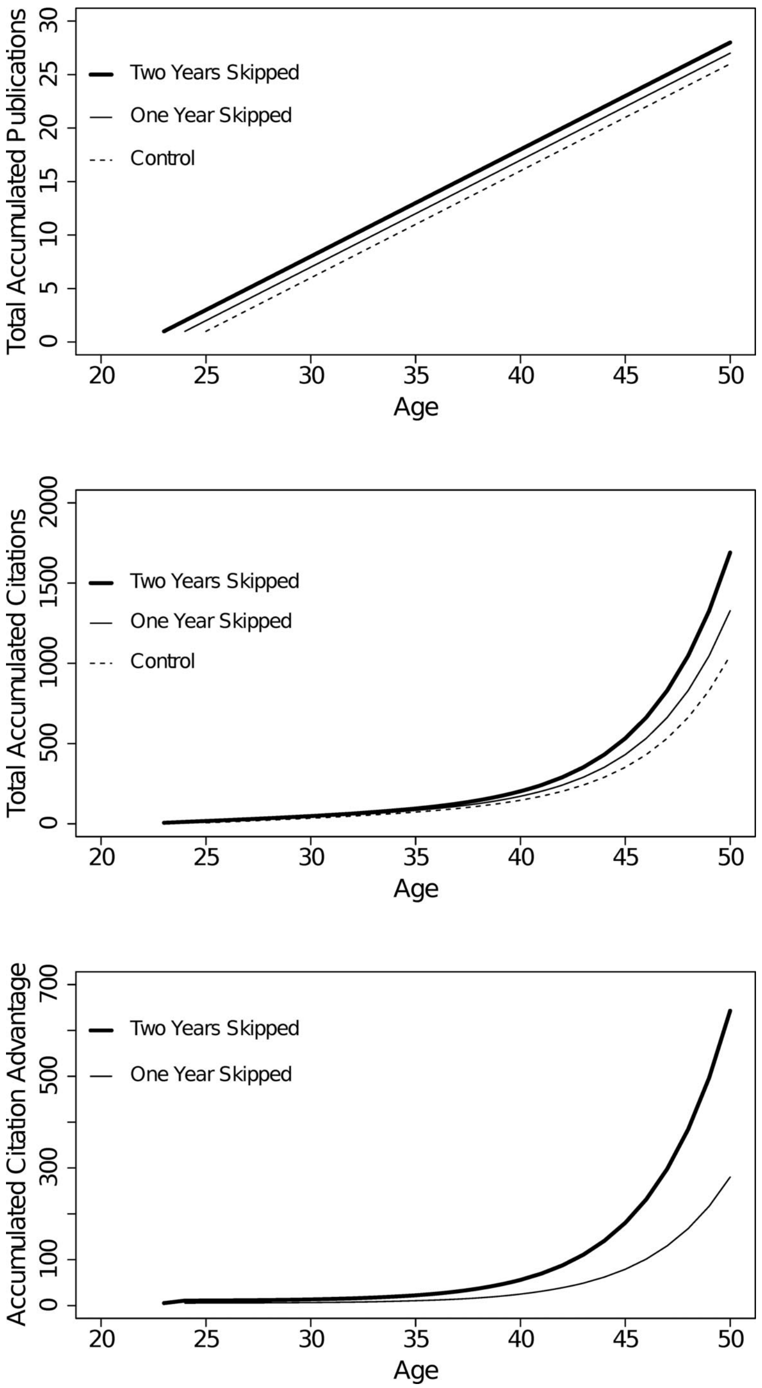

Figure C2. A hypothetical example of a small effect in initial starting points resulting in large differences in midcareer. The top panel shows the cumulative publications of the same individual, publishing one article per year, under three possible starting ages (24, 25, and 26 years). The middle panel shows their cumulative citations, where citations at age t = (5)(articles$_t$) + 1.3 articles$_t$, and articles$_t$ is the total number of published articles accumulated by age t. The bottom panel shows the slowly accumulated advantage in citation counts, granted by grade skipping, compared to no skipping at all.

(*Appendices continue*)

Appendix D

Pre-Imputation Means and Missingness of Baseline Measures

Tables D1, D2 and D3 present means on baseline measures for all controls and grade skipping participants prior to missing data imputation, and Tables D4, D5, and D6 present the number of observations and percentage of total observations missing on each of the baseline measures prior to missing data imputation.

Table D1
Pre-Imputation Baseline Means of 1972 Cohort

1972 cohort	Controls	Grade skippers
N	1,753	179
SAT Math	517	569
Mother's highest degree	3.2	3.7
Father's highest degree	4.3	4.5
Mother's occupational prestige	74	75
Father's occupational prestige	77	79
Birth order	2.1	2.0
Number of siblings	2.4	2.3
Liking for school	3.1	3.1
Liking for math	3.4	3.6
Doing well in math class	2.9	3.0
Learning math	1.3	1.4
Math importance	4.4	4.5
Previous grades skipped	0.1	0.2
Proportion male	0.62	0.57

Note. Means and proportions of 14 background variables measured at age 13 years across controls and grade skippers prior to missing data imputation. Liking, doing well, and importance variables refer to items presented to participants at their initial identification.

Table D2
Pre-Imputation Baseline Means of 1976 Cohort

1976 cohort	Controls	Grade skippers
N	507	116
SAT Math score	548	579
SAT Verbal score	456	486
Mother's highest degree	4.5	4.8
Father's highest degree	5.2	5.4
Number of siblings	1.8	1.7
Liking for school	3.9	3.9
Liking for math class	4.3	4.4
Liking for biology class	3.5	3.6
Liking for chemistry class	3.8	3.9
Liking for physics class	3.7	3.9
Doing well in math class	1.9	1.8
Doing well in science class	2.0	1.9
Learning math	1.3	1.6
Learning science	1.2	1.2
Math importance	3.6	3.6
Biology importance	2.6	2.4
Chemistry importance	2.8	2.8
Physics importance	2.9	3.2
Previous grade skipped	0.1	0.2
Proportion male	0.7	0.7
Proportion in public school	0.82	0.84

Note. Means and proportions of 21 background variables measured at age 13 years across controls and grade skippers prior to missing data imputation. Liking, doing well, and importance variables refer to items presented to participants at their initial identification.

(*Appendices continue*)

Table D3
Pre-Imputation Baseline Means of 1980 Cohort

1980 cohort	Controls	Grade skippers
N	167	68
SAT Math score	683	719
SAT Verbal score	549	557
Mother's highest degree	6.0	5.8
Father's highest degree	7.0	6.7
Mother's occupational prestige	69	69
Father's occupational prestige	80	81
Number of siblings	1.4	1.4
Liking for school	3.8	4.0
Liking for math class	4.6	4.8
Liking for biology class	3.7	3.8
Liking for chemistry class	4.1	4.3
Liking for physics class	4.2	4.4
Liking for English class	3.8	3.9
Liking for writing	3.5	3.4
Liking for foreign language class	4.1	4.0
Liking for social studies	3.6	3.8
Learning math	1.4	1.7
Previous grades skipped	0.5	0.4
Proportion male	0.74	0.93
Proportion in public school	0.79	0.85

Note. Means and proportions of 20 background variables measured at age 13 years across controls and grade skippers prior to missing data imputation. Liking, doing well, and importance variables refer to items presented to participants at their initial identification.

Table D4
Missingness in 1972 Cohort

1972 cohort	*N* missing	% missing
SAT Math	149	7.7
Mother's highest degree	162	8.4
Father's highest degree	161	8.3
Mother's occupational prestige	577	29.9
Father's occupational prestige	349	18.1
Birth order	99	5.1
Number of siblings	99	5.1
Liking for school	170	8.8
Liking for math class	513	26.6
Doing well in math class	519	26.9
Learning math	510	26.4
Math importance	572	29.6
Previous grades skipped	0	0.0
Sex	0	0.0

Note. Number and percentage of observations missing on covariates prior to missing data imputation in the 1972 cohort. Liking, doing well, and importance variables refer to items presented to participants at their initial identification.

(*Appendices continue*)

Table D5
Missingness in 1976 Cohort

1976 cohort	*N* missing	% missing
SAT Math	7	1.1
SAT Verbal	7	1.1
Mother’s highest degree	3	0.5
Father’s highest degree	7	1.1
Number of siblings	6	1.0
Liking for school	2	0.3
Liking for math class	9	1.4
Liking for biology class	68	10.9
Liking for chemistry class	106	17.0
Liking for physics class	121	19.4
Doing well in math class	2	0.3
Doing well in science class	45	7.2
Learning math	6	1.0
Learning science	40	6.4
Math importance	33	5.3
Biology importance	51	8.2
Chemistry importance	59	9.5
Physics importance	61	9.8
Previous grades skipped	0	0.0
Sex	0	0.0
School type	1	0.2

Note. Number and percentage of observations missing on covariates prior to missing data imputation in the 1976 cohort. Liking, doing well, and importance variables refer to items presented to participants at their initial identification.

Table D6
Missingness in 1980 Cohort

1980 cohort	*N* missing	% missing
SAT Math	1	0.4
SAT Verbal	1	0.4
Mother’s highest degree	9	3.8
Father’s highest degree	1	0.4
Mother’s occupational prestige	8	3.4
Father’s occupational prestige	8	3.4
Number of siblings	0	0.0
Liking for school	2	0.9
Liking for math class	1	0.4
Liking for biology class	4	1.7
Liking for chemistry class	4	1.7
Liking for physics class	4	1.7
Liking for English class	1	0.4
Liking for writing	2	0.9
Liking for foreign language class	1	0.4
Liking for social studies	2	0.9
Learning math	2	0.9
Previous grades skipped	24	10.2
Sex	0	0.0
School type	1	0.4

Note. Number and percentage of observations missing on covariates prior to missing data imputation in the 1980 cohort. Liking, doing well, and importance variables refer to items presented to participants at their initial identification.

Received May 24, 2011
Revision received June 11, 2012
Accepted June 15, 2012 ■