WHEN LESS IS MORE:

EFFECTS OF GRADE SKIPPING ON ADULT STEM ACCOMPLISHMENTS AMONG MATHEMATICALLY PRECOCIOUS ADOLESCENTS

By

Gregory J. Park

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

August, 2011

Nashville, Tennessee

Approved:

Professor David Lubinski

Professor Camilla P. Benbow

Professor James H. Stieger

Professor Andrew J. Tomarken

To my parents,
Steven and Anne,
for their love, sacrifices, and support.

ACKNOWLEDGMENTS

I am grateful to the faculty, friends, and family that made this work possible.

Support for this research was provided by a Research and Training Grant from the Templeton Foundation and the National Institute of Child Health and Development Grant P30HD15052 to the John F. Kennedy Center at Vanderbilt University.

My co-advisors, Dr. David Lubinski and Dr. Camilla P. Benbow, shared with me not only an invaluable collection of data but also their vision of what is psychologically substantive. I am grateful for their continued support throughout my doctoral training. I would also like to thank Dr. Julian C. Stanley for creating the Study of Mathematically Precocious Youth. Without him, none of my research would have been possible.

I thank my committee members, Dr. James Steiger and Dr. Andrew Tomarken, for their comments, suggestions, and ability to ask tough questions in a gentle way. Through their teaching and personal feedback, they greatly improved the quality of this dissertation.

I also thank my colleagues, fellow students, and friends, Stijn Smeets and Kim Ferriman Robertson, who shared their insights on my countless programming bugs, graphs, and early drafts.

Most importantly, I am grateful for the support and faith of my parents, who have always been my greatest advisors in all my pursuits, and for Julie, her love, patience, and continual inspiration.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Acceleration is a broadly defined educational intervention that takes many disparate forms in practice, but these forms generally fall into one of two groups, grade-based and subject-based (Southern & Jones, 2004). Grade-based acceleration, such as skipping one or more grades, shortens the total time spent in an educational program by allowing students to move to more developmentally appropriate content by skipping over what they already know or can easily and rapidly assimilate (Stanley, 2000). Alternatively, subject-based acceleration, also known as content-based acceleration, increases the progression through an educational program by increasing the rate, complexity, or depth of content, while keeping students with their same-age peers for most of the day. Examples of subject-based acceleration included Advanced Placement (AP) courses, college courses in high school, or summer enrichment programs. For methodological reasons explained below, the focus of the current study is grade-based acceleration, and unless otherwise noted, the terms "grade-based acceleration", "acceleration", and "grade skipping" will be used interchangeably.

Interest in the potential of acceleration to optimize talent development has a long history, spiking in the middle of the 20th century. Earlier, Seashore (1922) suggested that the most intellectually talented students, and in turn the arts and sciences, could benefit from increasing the rate at which they move through the educational system, yet the "gigantic educational machinery" at the time was unable to properly

accommodate such students. The onset of World War II intensified this interest, as the successful cultivation of technical and scientific skills and productivity became an issue of national security. Pressey (1946a, 1946b, 1955) argued that grade-based acceleration had potential to save valuable time during a critical period in precocious individual's development and suggested a theory of how acceleration may increase overall career productivity. Individuals, Pressey claimed, have a "prime" in early adulthood in which 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] Terman (1954), drawing on his own pioneering work with talented individuals and the research on age and achievement by Lehman (1946), 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).

The requirement of earning advanced training in scientific and technical fields often demands the student to be bogged down in training throughout these peak years, and this may "curtail maximum fruitfulness of a professional career" (Pressey, 1946a, p. 324). A proposed solution to this inefficiency was the grade-based acceleration of students based on their individual differences in rates of learning, freeing them of the usual lockstep, age-based educational track. By advancing more quickly through the educational system at high school and college age, the brightest students would lose little, if anything, but potentially gain intellectual development, interpersonal maturity, and most importantly, time. Work by contemporaries such as Paterson

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

(1957) and Hobbs (1951, 1958) further reinforced the idea that tailoring educational opportunities to individuality could reveal untapped resources, yielding benefits for both students and society.

Within the last decade, the rapid increase in globalization lead 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., American Competitiveness Initiative, 2006; Friedman, 2005; National Science Board, 2010a). Scientific and technological innovation are seen as drivers of national economic growth and quality of life improvement, and the identifying and development of domestic STEM talent is a national concern.

On a smaller scale, individuals looking to save time and institutions attempting to cut costs may find some relief in grade-based acceleration. The increased cost in time and money required to earn educational degrees has motivated institutions, such as the Northwestern University School of Law, to offer degrees on a compressed schedule (two years instead of three for a law degree; Mangan, 2008), while financial constraints have pushed some states to consider making optional the entire senior year of high school in an effort to reduce costs (Correll, 2010). The potential of properly applied acceleration to increase productivity, individual satisfaction and well-being, as well as time and money for the individual and institutions, is perhaps more pertinent now than at the time of Pressey's initial urging over sixty years ago.

Despite the many calls for increased application of grade-based acceleration, the longitudinal data necessary to support these suggestions empirically have been scarce. The current study uses data from a study of mathematically precocious individuals,

tracked for over 30 years, to investigate several hypotheses implied by the time-saving theory. Mathematical precocity is particularly relevant here, as mathematical ability has been demonstrated to be useful indicator of potential for STEM talent (Super & Bachrach, 1957; Lubinski & Benbow, 2006; Wai, Lubinski, & Benbow, 2009). After briefly elaborating on existing theory and literature on grade-based acceleration, the focus is shifted to handling the problems of selection bias that are common in observational data. Following this, a combination of matching techniques is described and proposed as a useful methodological strategy for testing hypotheses from the time-saving theory.

CHAPTER II

THE TIME-SAVING THEORY

In order to generate testable hypotheses about grade-based acceleration and its effects on STEM outcomes in adulthood, some additional elaboration of Pressey's general theory is necessary. This theory, referred to as the *time-saving theory*, is intuitively appealing but difficult to test. Figure 1 illustrates the major components of one interpretation of the theory with arrows indicating the direction of causal flow. While this is a simplification of the causal system, it outlines its major components and assumptions and motivates a strategy for handling the methodological obstacles common to research in educational acceleration.

According to the time-saving theory, grade-based acceleration has a direct effect on a precocious individual's career choices in two ways. First, by decreasing the time spent in primary and secondary school, those who have accelerated (*grade skippers*) are likely to enter and finish undergraduate programs earlier and will be presented with the same educational and occupational choices of a precocious non-skipper but at an earlier age. By being slightly earlier in the "prime", grade skippers may be more likely to pursue additional training in graduate or professional school. For example, a 21 year old recent college graduate may be slightly more likely to decide to enter a five year PhD program than a 22 year old recent graduate.

Secondly, grade skippers will finish this additional training and transition into

their careers slightly earlier than their non-skipping peers. At the end of their respective graduate programs, the grade skipper is 26 and the non-skipper is 27. The relationship between age at career onset and adult productivity, particularly in STEM fields, has been the focus of several researchers throughout the last century (Lehman, 1946, 1953; Dennis, 1956; Zuckerman, 1977; Simonton, 1988, 1997). A consistent finding is an individual's productivity over time is usually captured fairly well by a single-peaked curve, but individual curves vary in their onset, the rate of acceleration, peak height, and rate of deceleration. A second finding is that early productivity is associated with greater productivity across the rest of the lifespan, and this finding has important implications in the context of the time-saving theory. If career onset has no effect on productivity and is only correlated with other determinants of productivity, then shifting an individual's productivity curve towards an earlier onset will not affect total productivity. In that case, early career onset would lead to an equally earlier career termination. However, the time-saving theory suggests that earlier career onset affects the *shape* of the productivity curve in some way, perhaps by increasing the rate of acceleration towards a career peak or by stretching the entire curve out, increasing the total time of the career.

This effect on overall career productivity is the third, and most important component of the time-saving theory. It states that small age differences benefit grade skippers throughout their careers by allowing an extra year of this prime to be freed from training, and this small edge results in cumulative effects realized over the course of their careers. Part of this benefit from acceleration is mediated (Baron & Kenny, 1986) by the effect on the age at career onset. This effect is captured in the

path *ab*, or Acceleration → Career Onset → Productivity. Acceleration may have additional effects on adult productivity that are unmediated by the time advantage and have other causal paths, captured by path *c*. For example, grade skipping in high school may increase an individual's confidence or interest in school, and these effects influence productivity regardless of the age advantage.

Some longitudinal studies have followed up accelerated students for 10 years or more, finding evidence indirectly supporting the notion that acceleration affects the time of career onset, by decreasing the age at which students finish their formal education. Flesher and Pressey (1955) found that women who participated in an accelerated college program during the 1940s not only completed their four-year degree in three years or less, but were also more likely to pursue and earn graduate degrees. Pressey (1967) followed students ten years after their initial acceleration between 1951 and 1954, and many had entered college at least one year early. Compared to their peers, they earned slightly more graduate degrees and, on average, earned them two years earlier. In adulthood, several of these participants indicated that decreasing the time spent in high school had a positive effect on their well-being and that entering their professional field at a younger age came with considerable advantages. More recent findings come from the Study of Mathematically Precocious Youth (SMPY; Lubinski & Benbow, 2006). The study's founder, Stanley (1973), noted that many students in the sample were more than able to skip one or more grades or enter college early. Following up on this, Swiatek and Benbow (1991) found that mathematically

precocious students who entered college at least one year early had higher educational attainment and tended to enter graduate school about one year earlier than their ability-matched, non-accelerated peers.

A recent study by Wai, Lubinski, Benbow, and Steiger (2010) focused on long-term effects of a combination of grade-based acceleration, subject-based acceleration and enrichment opportunities within mathematically precocious students. Participants who received a high dose of these accelerative opportunities were more likely to earn science, technology, engineering and mathematics (STEM) PhDs, author a STEM peer-reviewed publication, earn a patent, and secure a tenure-track position in a STEM field. These findings are in line with predictions from the time-saving theory, but the correlations between academic interests, cognitive abilities, and accelerative dosages in this sample also complicate inferences about the unique contribution of acceleration.

Returning Figure 1, one can see the confounding effect of ability throughout the system. Students are often accelerated based on their cognitive ability, although other personal attributes are important for successful implementation and are always taken into account for best practices (Colangelo, Assouline, & Lupkowski-Shoplik, 2004). Because ability has an effect on career onset and adult productivity while also affecting acceleration, it acts as a confounding variable. To see this in Figure 1, note that ability can affect career onset directly via path e or through acceleration via path da. Similarly, ability can affect productivity through four paths: directly (f), through career onset (eb), through acceleration (dc), and through both acceleration and career onset (dab). The result is that studies of the time-saving theory that only

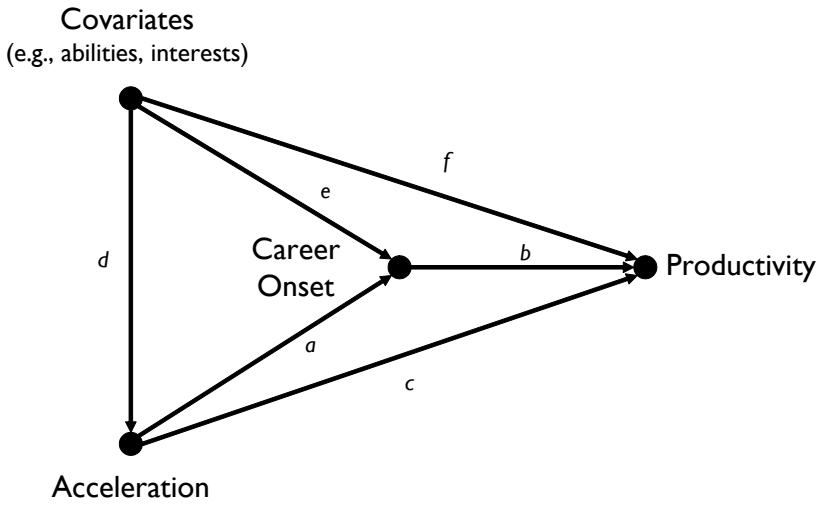

Figure 1: The basic components of the time-saving theory. Single-headed arrows represent unidirectional causal flow.

focus on paths a, b, and c will be biased by the confounding influence of ability and other attributes contributing to the assignment of acceleration.

The ideal methodological solution to this problem is randomization of the accelerative treatment, which would effectively break or block the back-door path d. If this path is blocked, estimates of path a and c are unbiased estimates of the effect of acceleration with respect to ability. However, even if randomization of acceleration were possible, interpretation and estimation of path b, the effect of age at career onset on later productivity, is complicated for several reasons general to mediational analyses (Cole & Maxwell, 2003; Bullock, Green, & Ha, 2010). First, acceleration may affect unobserved variables that influence both career onset and later productivity. Secondly, the effect of career onset on productivity may differ across individuals based on a third variable, also known as moderated mediation. For example, males and

females may respond differently to finishing an advanced degree earlier than usual. However, mediation may be moderated by one or more unobserved variables, as well, creating misleading estimates of the mediated effect. Thirdly, there may be heterogeneity in the effect of acceleration on career onset. Given that these problems exist in the optimal case of a randomized experiment, enthusiasm for uncovering mediation in observational studies should be tempered, even with large samples and reliable measures.

The difficulties in the studies of long-term effects of acceleration are common throughout much of social science and policy research. The lack of randomized treatment assignments in studies of acceleration complicates possible inference, a problem reflected in a recent report of the National Mathematics Advisory Panel (2008). Without randomization, accelerated and control groups are almost guaranteed to differ systematically in their distributions of important variables, which are likely related to outcomes of interest in the study. However, in studies of acceleration, like education and much of the rest of the social sciences, most data are available only as the result of observational studies. Lack of randomization in social science research is often due to insurmountable logistical and financial obstacles or ethical considerations (Rubin, 1974; Shadish, Cook, & Campbell, 2002), and research on acceleration is no exception. If a student or his or her parents desire acceleration, they are unlikely to wait for random assignment if they can simply find opportunities through other means. The panel categorized the majority of the existing research on acceleration, which is largely observational and nonrandomized in nature, as being of moderate or low quality (National Mathematics Advisory Panel, 2008, p. 155). While the final

report did find positive effects of acceleration from the few qualifying studies, the potential estimation error from most of the quasi-experimental studies was considered to be too overwhelming.

Dealing with estimation error is one of the major tasks of researchers working with observational data, and a clear understanding of its sources should guide analysis. While no longitudinal data from studies incorporating randomized assignment of acceleration exist, randomization is just one of many design features used to cope with estimation error, and inferences from observational studies can be greatly improved by incorporating appropriate design features, such as matching.

CHAPTER III

MATCHING

The current study uses matching to reduce bias in effect estimates resulting from covariate imbalance. The traditional justification for matching comes from a formalization of the process of causal inference, known as the potential outcomes framework (POF) or Neyman-Rubin Causal Model (Rubin, 2010). A brief overview of the major elements of the POF is an important step in establishing the validity of the matching procedures used in the current study. The added benefit of reduced model dependence is also discussed. Finally, the language of the POF is used to clarify some of the controversy around the application of matching that exists in the psychological literature.

Causal Inference

The POF, sometimes referred to as the potential outcomes model or the Neyman-Rubin Causal Model, has historical roots in Jerzey Neyman's Ph.D. thesis (Splawa-Neyman, 1923/1990), in which potential outcomes notation was first used (Rubin, 2005). This approach was further elaborated through a series of articles by Rubin and colleagues (e.g., Rosenbaum & Rubin, 1983; Rubin, 1974, 1986, 2010; Shadish, 2010). The idea of potential outcomes (or "counterfactuals") is at the heart of the POF. For current purposes, assume that a binary cause or treatment is the focus

of a given study. Corresponding to this binary treatment are two possible states of existence for every member of a population, treatment or control. For example, if a skipping an entire grade is the treatment, it is assumed to be binary. A subject either skips a grade (the treatment state) or not (the control state). In this study, the outcome of interest is y, and this outcome for an individual i is y_i. Individual i has two potential outcomes, depending on whether he or she is assigned to the treatment or control. If i is assigned to the treatment, we will observe y_i^1. If i is assigned to the control, we will observe y_i^0. The individual-level treatment effect, TE_i, is simply

$$\text{TE}_i = y_i^1 - y_i^0. \tag{1}$$

The obvious problem is that only one of these outcomes is actually observable for i, otherwise known as the "fundamental problem of causal inference" (Holland, 1986). If the subject i skips a grade and y_i^1 is observed, it is impossible to observe y_i^0, or what *would have happened* had i not skipped a grade (the counterfactual).

Consider that a sample of n individuals are drawn from a population N. Every individual has a treatment assignment T_i, where i receives the treatment if $T_i = 1$ or control if $T_i = 0$. Y is a random variable and the outcome of interest. Although the treatment effect TE_i is never directly observable, the average treatment effect (ATE) is often estimated by the difference of the expected value of Y between individuals in the treatment (Y^1) and control groups (Y^0),

$$\text{ATE} = E(Y^1) - E(Y^0), \tag{2}$$

which is often simply the difference in means between the treatment and control groups. For binary outcomes, common in epidemiological research, the quantity of interest may be the risk ratio,

$$\frac{\pi_1}{\pi_0}. \tag{3}$$

where π_1 is the proportion of event occurrence in the treatment group and π_0 is the proportion of event occurrence in the control group. Therefore aggregated information from the treatment and control groups allows estimation of the unobservable quantity of interest, the ATE across the entire population. However, the quality of this estimate lies on several assumptions, many of which are untestable and unlikely to hold (Morgan & Winship, 2007).

Fewer assumptions are made when the focus is a subquantity of the ATE, the average treatment effect on the treated (ATT), interpreted as the average treatment effect for those who are likely to receive the treatment. The ATT is usually the quantity of interest in studies using matching, wherein control units are dropped or replicated until the control sample is balanced with the treatment sample. The treatment sample and their covariate distributions are used as a baseline, and the control group is constructed in order to create a sample as similar as possible to the treatment group with one exception: they did not receive the treatment. In the context of grade-based acceleration, the ATT can be interpreted as the average effect of the acceleration on the participants most likely to have been accelerated.

The Assignment Mechanism and Ignorability

Within the framework of the POF, there is a distinction between the data and the assignment mechanism, or the process that generates the observed outcomes and treatment assignment (Rubin, 1991). The assignment mechanism is the process that assigns a value of T to each unit and is formalized in the probabilistic statement

$$\Pr(T|X, U, Y^1, Y^0) \tag{4}$$

where T is a vector of treatment assignments, X is a vector of an observed covariate, U is a vector of an unobserved covariate, and Y^1 and Y^0 are potential outcomes. In other words, the assignment mechanism relates the treatment assignment T to observed and unobserved covariates and potential outcomes. Potential outcomes Y^1 and Y^0 are not known and the statement is usually simplified to

$$\Pr(T|X, U). \tag{5}$$

or the probability of T given X and U. This generalizes to cases with many observed covariates X and unobserved covariates U. The goal of many design features is to break the dependence between T and covariates X and U, with the randomization of the treatment assignment often being held as the gold standard (Campbell & Stanley, 1963; Shadish et al., 2002). Randomizing the treatment assignment makes the assignment mechanism random, breaking the dependence on X and U and justifying

the assumption that

$$Y^1, Y^0 \perp T \tag{6}$$

or that the distribution of potential outcomes is ignorable with respect to the treatment assignment. When randomization is not possible, methods such as statistical adjustment or matching aim to justify the assumption

$$Y^1, Y^0 \perp T \mid X, U, \tag{7}$$

the distribution of potential outcomes is ignorable with respect to the treatment assignment given X and U (Gelman & Hill, 2007). When the treatment assignment is non-ignorable, imbalance between the treatment and control groups on X and U creates systematic error in the estimation of the treatment effect. Matching is an attempt to reduce or eliminate components of this error.

The Stable Unit Treatment Value Assumption

A key assumption of the POF is the stable unit treatment value assumption (SUTVA; Rubin, 1980b, 1986, 2010), which holds that the value or amount of the treatment is stable across units. SUTVA can be broken into two smaller components. The first, which may be called the *hidden treatment component*, states that there are no versions of the treatment hidden from the analysis. In the grade skipping example, if the treatment is considered binary, but some subjects in the treatment group skip one year while others skip two, SUTVA is likely violated. Whether it is or

not depends on the nature of the outcome Y. Imagine that a study aims to estimate gains on an achievement test resulting from grade skipping. If Y is an achievement test score, and different amounts of grade skipping actually do result in differing levels of improvement on this score, then SUTVA does not hold. The treatment value is no longer stable across treated indivduals, because it depends on how many grades they skipped, and this is not apparent given the binary nature of the coding. However, if the study has a slightly weaker aim, to determine whether grade skipping gives *any* gains at all, then SUTVA *does* hold even if varying levels of grade skipping result in varying sizes of improvement.

The second component of SUTVA may be called the *no interference component*. This aspect concerns the relationship between the treatment assignment mechanism and the treatment value and holds only if the treatment value is independent of the actual assignment mechanism. Rubin (2010) explains this in the context of a job training program. The treatment effect of the training program must be stable across all possible assignments. This means that the effect is independent of not only who receives it but *how many* receive it. For example, the training program may be highly effective when given to only 5% of the local workforce but highly ineffective if given 50%, due to flooding the local labor market with these improved skills.

SUTVA has important implications for research on the effects of acceleration in observational studies. Because acceleration is a very loosely defined intervention, it can take many different forms (Wai et al., 2010). Further complications result from the fact that the same type of acceleration can greatly vary between schools, regions, and points in time. The Advanced Placement (AP) system is a good example of such

a complication. In SMPY, approximately 25% of the 1972-1974 cohort enrolled in AP courses compared to almost every participant in the 1980-1983 cohort (Wai et al., 2010). One reason for this may be that the 1980-1983 cohort was more able and motivated, and therefore enrolled in AP courses at a higher rate. However, there has been a gradual increase in the availability and accessibility of AP courses since their introduction. If increased access to AP courses also changed the actual coursework in the process (e.g. making them less demanding), SUTVA is violated, and the treatment is not stable across participants. There is evidence suggesting that this is indeed occuring (Lichten, 2000; Lewin, 2002; Bleske-Rechek, Lubinski, & Benbow, 2004). Even if analyses were restricted within each cohort, it is reasonable to assume that different schools offer different versions of the same AP course. Extending this reasoning to other forms of acceleration, such as advanced subject matter placement, summer coursework, or college courses in high school, makes the analysis of the effect of subject-based acceleration a very difficult enterprise outside of a tightly controlled experimental framework.

Grade-based acceleration, such as grade-skipping, fares much better. It has been handled more conservatively than some types of subject-based acceleration for many reasons (Colangelo, Assouline, & Lupkowski-Shoplik, 2004), and it has not been subject to the same widespread increase in accessibility or application. It is also easier to quantify: the treatment value is the amount of time saved, and there is a clear distinction between different levels of grade-based acceleration (i.e., one, two, or three years of skipping are discrete levels of treatment). Furthermore, more stability in the treatment value is secured by limiting the actual grade level that is skipped

(only comparing subjects who skip grades in high school, for example). For these reasons, grade-based acceleration holds the most promise as a treatment variable in the framework of the POF.

Matching to Reduce Estimation Error

When the treatment in an observational study is well-defined, SUTVA holds, and relevant pre-treatment covariates are observed, matching can improve causal inference, primarily by reducing estimation error. Estimation error is defined as the difference between the population quantity of interest, in this case the population average treatment effect on the treated (PATT), and the estimated effect, D, from the sample data. Borrowing from Imai, King, and Stuart (2008), this difference is called Δ. This can be decomposed into *sampling selection error*, Δ_S, or error arising from the sampling process, and *treatment imbalance error*, Δ_T, or error arising from the imbalance of observed covariates X and unobserved covariates U across treatment and control groups.

These can be further decomposed with respect to X and U. Δ_S is composed of Δ_{S_X}, sampling selection error with respect to observed covariates X, and Δ_{S_U}, sampling selection error with respect to unobserved covariates U. Likewise, Δ_T is composed of Δ_{T_X} and Δ_{T_U}.

The ideal study uses a random sample from the population of interest and randomly assigns the treatment. As sample size increases, each component of estimation error approaches zero. In an observational study of a special population, the sample is not drawn randomly and the treatment assignment is not randomized, so Δ_S and

Δ_T are a cause for concern. Researchers can avoid the problem of Δ_S completely by switching the quantity of interest from the PATT to the sample average treatment effect on the treated (SATT). In other words, inferences based the estimated quantities in the study should only be applied to the sample. External validity of estimates may be established by comparing several studies of the similar samples through replication, if necessary.

Even if one is temporarily willing to assume the SATT as the quantity of interest, Δ_T is still a problematic source of error in observational studies. Treatment and control groups will usually be imbalanced on both observed and unobserved covariates, and ignorability cannot be assumed. The goal of most design features, such as matching or blocking, is balancing groups on observable and observed covariates, thereby reducing or eliminating Δ_{T_X}. Matching is the method of choice when treatment assignment has already occurred and *pre-treatment* covariates are observed.

In practice, matching is done by removing or replicating observations in the control group until both the treatment and control groups are balanced on all observed pre-treatment covariates. Once balance is achieved, the treatment assignment can be assumed ignorable with respect to the observed covariates. However, the problem of imbalance due to unobserved covariates, Δ_{T_U} still remains, but if X and U are related, the quantity changes to $\Delta_{T_{U|X}}$. This remaining error can be handled in a number of ways. In some cases, it may be assumed that there are no unobserved variables, and $\Delta_{T_{U|X}}$ is assumed to be equal to zero. A more realistic approach is to ensure that the observed variables X are the most important to the assignment mechanism and hope that the remaining unobserved variables U are correlated with

X. Thus, by matching on the critical variables X, the assignment mechanism is reasonably estimated, imbalance on U is reduced to a trivial level. Recent empirical research, comparing the effect estimates from randomized experiments to those from observational studies with matching, has demonstrated the utility of matching in reducing bias, provided that reliable and relevant covariates are incorporated into the matching procedure (Shadish, Clark, & Steiner, 2008; Cook & Steiner, 2010; Steiner, Cook, & Shadish, 2011).

In the current study, the observed variables X should be related to both the propensity to skip grades and important scientific outcomes in adulthood. The background variables available for matching are promising, including measures of quantitative and verbal ability, preferences for various academic subjects, self-perceived ability in math and science, and demographic variables such as family size, parental education attainment, and parental occupational prestige. If matching successfully breaks the back-door path between X and T, then estimates of effects are now based on a causal system represented by Figure 2. Provided that only pre-treatment covariates are used in matching and the subsequent analysis, the differences on various outcomes between the grade skippers and their matched controls are unbiased (or, perhaps more realistically, much less biased) estimates of the effects of grade skipping.

Reducing Model Dependence

Matching is often justified as a means of improving causal inference, but this also relies heavily on the assumption of ignorability, which may be dubious and unverifiable. However, (Ho, Imai, King, & Stuart, 2007) demonstrated how matching can be a powerful method of reducing the model dependence of effect estimates. All matching methods work by dropping, replicating, or weighting observations prior to any analysis. This *preprocessing* step results in a reduced number of total observations, but now treated and control observations will have much greater overlap in their covariate distributions.

If the next step in the analysis uses, for example, a regression model to estimate the treatment effect while simultaneously controlling for other covariates, the estimated treatment effect (usually a regression coefficient) will be stable over various specifications of the regression model. This benefits the analyst by reducing the uncertainty inherent in model selection, but it also reduces the potential for "bad" behavior on the analyst's part. Because the treatment effect estimate is much more resistant to model specification after matching, the analyst's ability to hunt for the model with the most favorable estimates is greatly reduced.

In the current study, logistic regression is used to estimate the effect of grade skipping while controlling for all observed pre-treatment covariates. The reduction of model dependence is particularly important in this case, because between 14 and 22 pre-treatment covariates are observed in each analysis. How to choose the "best"

logistic regression model specification from the model space is a classic problem, but here it is mitigated largely by preprocessing of the data to create balanced groups.

Matching Methods

Methods of matching are evolving rapidly (Ho et al., 2007; Diamond & Sekhon, 2006; Sekhon, 2009, 2007; Imai et al., 2008; S. M. Iacus, King, & Porro, in press). The oldest method is known as *exact* matching, in which control observations are matched with treatment observations with exactly the same values on X. This is highly effective when only one or two covariates are of interest, but exact matching quickly becomes infeasible when X is large due to the curse of dimensionality (Bellman, 1961). In most cases, groups are imbalanced on many covariates, so reducing the dimensionality of X is necessary for matching in most samples.

Propensity score matching (Rosenbaum & Rubin, 1983) counters the high dimensionality of X by reducing any individual observation of X to a single value between 0 and 1, 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 X. Units are then matched based on the propensity score, most commonly using nearest-neighbor matching. This method starts with a given treatment unit and its propensity score and then selects the control unit with the closest propensity score. 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. An alternative to propensity score matching, Mahalanobis distance matching, matches each treatment unit with one or more control units based on the Mahalanobis distance between them (Rubin, 1980a). These methods can also be combined, by exactly matching observations on an important covariate and then to the nearest neighbor on the propensity score, for example.

A common problem can arise when either propensity score or Mahalanobis distance matching is used in practice. The ambiguity in this approach lies in the specification of the propensity score model, usually estimated using a logistic regression model in which the treatment assignment is the dependent variable and X are the inputs. The choice of specification of the model (i.e. whether to use transformations or include interactions) is up to the analyst, and the accepted practice is to use the specification that results in the best balance (Ho et al., 2007; S. M. Iacus et al., in press; S. Iacus, King, & Porro, 2011). The first problem is one of model choice. If X is reasonably large (10 or more variables), there are hundreds of thousands of possible specifications, so a manual search through all possible specifications can be difficult. The second problem is deciding how to assess balance.

One common approach is to compare the newly created treatment and control groups on each covariate using hypothesis tests, such as a t-test. If the null hypothesis of equal means is not rejected by the t-test, balance has been achieved. This approach has received strong criticism recently and its discouraged for several reasons (for a summary of these criticisms, see Imai et al., 2008), mainly because it encourages the deletion of control units and decreasing statistical power. If enough control units are deleted, even randomly, a t-test will never reject the null hypothesis.

An alternative approach is to abandon hypothesis tests and to simply use reasonable heuristics when comparing means. For example, if the difference between the means on a covariate is less than .25 (or .1) standard deviations, balance is achieved (Ho et al., 2007; Cochran, 1968). If not, respecify the model and reassess balance. However, only comparing means may miss important distributional features, such as discrepancies in the shape of distributions. The shape can be visually inspected with empirical quantile-quantile plots, and if too much discrepancy is observed, the analyst must respecify the model and try again. Obviously, this iterative process could continue endlessly without clear stopping criteria.

Genetic algorithms offer a solution to the ambiguity surrounding stopping criteria (Diamond & Sekhon, 2006; Sekhon, 2007). An example is the GenMatch algorithm, which combines propensity score with Mahalanobis distance score matching and automates the specification and balance assessment. A initial set of weights is chosen for the model, balance is assessed, the weights are changed, balance is reassessed, and this process continues until an optimal set of weights is chosen that maximizes balance, both in means and distributional shape of the covariates, between the two groups. The balance metrics for the algorithm are the p-values from t-tests of covariate means and two-sample Kolmogorov-Smirnov[1] tests. However, the desired ratio of control to treatment units is specified in advance by the analyst, so the rampant deletion of control units mentioned earlier is not an issue.

[1]The two-sample Kolmogorov-Smirnov test tests the hypothesis that two samples are drawn from the same sample based on the distance between their empirical cumulative distributions. Using this test is analogous to a visual inspection of the empirical quantile-quantile plots, with the advantage that it generates a test-statistic that can be minimized during by the iterations of the GenMatch algorithm.

Given the advantages of automating the model specification and matching process, the GenMatch algorithm is an attractive option for creating matched samples. However, propensity score matching is more mature, having been used in practice for several decades. It has a well-developed literature exploring its properties through simulation (e.g. Austin, 2010) and practice (Dehejia & Wahba, 1999; Shadish et al., 2008; Cook & Steiner, 2010; Steiner et al., 2011) and is arguably more transparent and intuitive than highly automated procedures, although such methods are quickly gaining popularity.

The Matching Fallacy

Meehl (1970, 1971) warned of three potential dangers of matching, and these criticisms have echoed throughout the psychological literature as the *matching fallacy* (e.g. Stigler & Miller, 1993; Kremen et al., 1996; Voglmaier et al., 2000). The alleged fallacy of matching is rooted in three unintended possible outcomes of matching: (a) matching on one covariate unmatches subjects on other covariates; (b) matching creates two samples that are unrepresentative of their respective populations; and (c) causal directionality is unclear and problems of causal inference arise.

Issue 1: Systematic Unmatching

The first issue, that matching will systematically unmatch subjects on some other covariate, is certainly an issue when exact matching is used with a small amount of subjects. To use the example from Meehl (1970), matching high school dropouts

with high school finishers on a measure of intelligence tends to create "systematic unmatching" on some other variable correlated with intelligence, such as need for achievement (*n Ach*). Comparing these matched samples on some later outcome, such as adult income, may be misleading if the treatment and outcome are functions of the unobserved covariate *n Ach*. Meehl contends that the dropout with a high IQ will be an extreme (low) deviate on *n Ach*, and will be matched with a high IQ graduate with the usual corresponding *n Ach*. On average, Meehl said and others repeated, this creates two samples matched on IQ but systematically unmatched on *n Ach*.

This problem is the result of exactly matching on only one covariate and ignoring another covariate that is important to the treatment assignment. There will always be unobserved variables that influence treatment assignment in observational studies, but a well-designed study will measure any variables that are known to be important to the assignment mechanism. When prior knowledge about the assignment is used to choose covariates that are correlated with potentially important unobserved covariates, the matching will reduce, not increase, the imbalance on the unobserved variables.

Issue 2: Creating Unrepresentative Samples

The second unintended consequence of matching occurs when exact matching on one or two variables creates samples that are highly unrepresentative of their respective populations. Meehl (1970) gives the example of matching schizophrenic individuals with manic-depressive peers on socioeconomic class. Supposed that the quantity of interest is the difference in population means on a measure of visual acuity between

individuals diagnosed with schizophrenia and those diagnosed with manic depression. In order to reduce the potentially confounding effect of socioeconomic class, the researcher matches subsets of each on socioeconomic class. However, the unintended result is that the schizophrenic sample is uncharacteristically high on socioeconomic class, while the manic-depressive sample is uncharacteristically low on socioeconomic class. While the quantity of interest is the difference in population means, the quantity that will now be estimated from the sample is quite different. The difference in sample means is now the difference between two unrepresentative samples, and it is unlikely to be useful in making generalizations to either the schizophrenic or the manic-depressive population.

However, unrepresentative samples can be useful in well-defined designs. In the case of acceleration, as in the current study, accelerated students are matched with control students on quantitative ability. Accelerated students are also higher on average than the control students on this variable, so the resulting matched control sample is uncharacteristically higher on quantitative ability than the control population. This is precisely the reason researchers switch the quantity of interest from the ATE to ATT. When the treatment is well-defined, the goal of matching is to find controls that are as similar as possible to the treated sample with the exception of treatment assignment.

Issue 3: Causal Ambiguity

The third criticism concerns ambiguity of causal directionality. Meehl (1970) gives the example of comparing subjects from different ethnic or religious backgrounds

(black and white subjects or Protestants and Catholics) on a variable such as IQ after matching on a third variable, such as socioeconomic class. Matching is done in this case to control for the effect of socioeconomic class on IQ. Causal inferences made after this matching are not possible, due to the causal effect of IQ on socioeconomic class or an underlying third variable causing both. Matching reduces a valid difference between the groups, an example of the "partialling fallacy" (Gordon, 1967, 1968).

This ambiguity is not a result of matching in itself but an ill-defined question due to the use of ethnic background as a treatment variable. In the POF, the treatment effect reflects the difference between two potential outcomes. Accordingly, the researcher must assume that potential outcomes are actually possible, even if only one is actually observed. In this case, it is unclear how the same individual could have the two potential outcomes from Meehl's example (as if same individual *potentially* could be either black or white, or either Protestant or Catholic). While mean differences between these two samples can be estimated, causal interpretations of this difference are not possible in the POF. Remaining mindful of potential outcomes is useful when determining whether matching is appropriate, and in this case, it is not.

Within the POF, several conditions must be met before causal inference is possible. The treatment must be well-defined, take place within a given time frame, and allow a reasonable acceptance of SUTVA. Additionally, the treatment assignment must be unconfounded with covariates in the treatment and control groups, and there must be overlap on X between the treatment and control groups. Matching often goes awry when these elements are ignored. By satisfying the usual conditions of the POF,

the unintended outcomes of matching are avoided. Under most conditions, matching will facilitate causal inference and improve estimates.

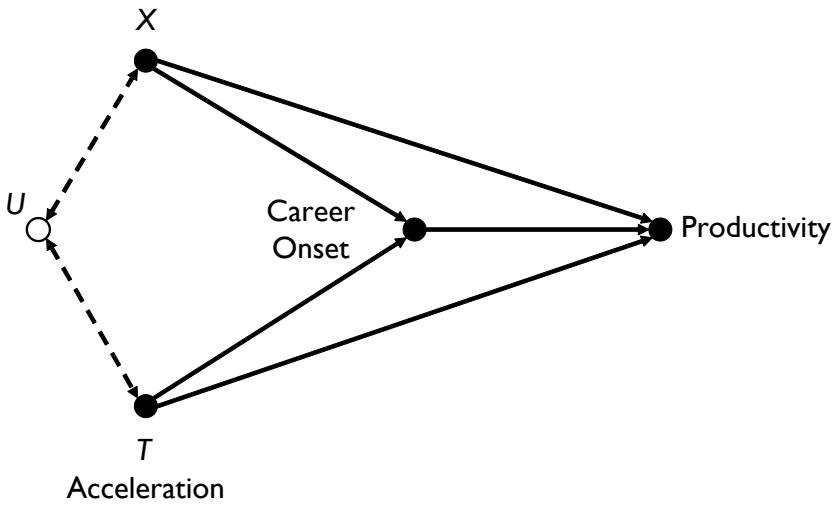

Figure 2: The basic components of the time-saving theory and assumptions of the current framework. X represents observed background variables used to create matched samples. U represents unobserved background variables. Dashed double-headed arrows represent correlations.

CHAPTER IV

METHODS

Design

The current study investigates three key hypotheses drawn from the time-saving theory, which predicts that grade-based acceleration can have an effect on career productivity in STEM fields. Using data from an observational study of thousands of mathematically precocious adolescents tracked longitudinally, the study aims to determine if those using grade-based acceleration, or grade skipping, *during high school*[1] (1) were more likely to pursue and earn advanced educational degrees and accomplishments in STEM fields, (2) if they reached these outcomes earlier than their non-accelerated, intellectual peers, and (3) if accelerated participants were more productive than non-accelerates when assessed at mid-career.

Using participants identified as mathematically precocious in adolescence has considerable advantages in 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 (Wai et al., 2009; Flanagan et al., 1962) and in samples of mathematically precocious individuals (Park, Lubinski, & Benbow, 2007, 2008; Lubinski & Benbow, 2006). Many of the regional and national indicators of

[1]Grade-skipping is restricted to high school to ensure that participants are matched on variables measured prior to the grade-skipping. One of these variables includes the number of grades skipped *prior* to identification.

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. For example, approximately 1% of the population is credited as an inventor on a patent (Huber, 1999). Population prevalence of STEM graduate degrees and publications are not known, but it reasonable to believe that they are similarly low.

Restricting the study to mathematically precocious individuals has the benefit of amplifying the baseline prevalence of these STEM outcomes, relative to the prevalence in a representative sample. Because the goal of this study is to estimate the average treatment effect on the treated, or the effect of grade skipping *among mathematically precocious individuals most likely to grade skip*, interest is in the shift in prevalence across the matched control and treated groups and not the prevalence in either subgroup alone.

Sample

Participants are 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, 1974 cohort, and 1980 cohort, respectively. Participants in every cohort were identified at or before age 13 by scoring in the top 1% of their age group on subtests of the College Board Scholastic Assessment Test (SAT). 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.

Each cohort was identified according to different cutpoints on the SAT subtests, described below. Using college entrance exams such as the SAT, which is a test designed for college bound high school students, provides additional ceiling room for such precocious students at age 13 and captures a greater range of their variation in their abilities than a standardized test designed for their age group. The 1972 cohort includes 2,188 participants 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) before age 13. These scores were estimated to be the lower bounds of scores of the top 1% of this age group, and many participants in the cohort scored well beyond this cutoff. The 1976 cohort includes 778 participants from Mid-Atlantic states scoring at least 500 on the SAT-Math or 430 on the SAT-Verbal before age 13, and these scores were estimated to be the lower bounds of scores of the top 0.5% of this age group. The 1980 cohort data contains information on 501 participants scoring at least 700 on the SAT-Math subtest or 630 on the SAT-Verbal subtest at or before age 13. These scores were estimated to be the lower bounds of scores of the top 0.01% of this age group.

After initial identification at or before age 13, participants were followed up at ages 18, 23, and 33 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 using public internet databases such as ProQuest Disseration and Thesis 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 in each cohort completed questionnaires including items about their academic preferences, perceived ability, number of siblings, and their parents' education and occupations, and these measures are used in the matching process. Several identical items were presented to participants in every cohort, and many typical items are listed in Table 1. For example, several items were in the form of "What word best describes your liking for X?" (where X may be math, physics, or English class) with potential responses ranging on a 5-point scale from "Strongly unfavorable" to "Strongly favorable". Items available for each cohort are given in the brief summaries below.

1972 Cohort

Table 2 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, and only SAT-Math scores were used in this study for this cohort. Participants listed the highest educational degree earned by each parent and their occupation. Highest educational degree was coded on a 1-7 scale (ranging from "less than high school" to "doctoral degree"), and the prestige of occupation was coded according to the socioeconomic index (SEI) from

Variable Name	**Item Description**	**Minimum**	**Maximum**
Mother's highest degree (1972)	Ordinal scale of highest degree earned	1 (Less than high school)	7 (Doctoral degree)
Father's highest degree (1972)	Ordinal scale of highest degree earned	1 (Less than high school)	7 (Doctoral degree)
Mother's highest degree (1976 & 1980)	Ordinal scale of highest degree earned	1 (Less than high school)	9 (Post-doctoral experience)
Father's highest degree (1976 & 1980)	Ordinal scale of highest degree earned	1 (Less than high school)	9 (Post-doctoral experience)
Mother's occupational prestige (1972 & 1980)	Occupational prestige according to Duncan (1961)	1 (low prestige)	100 (high prestige)
Father's occupational prestige (1972 & 1980)	Occupational prestige according to Duncan (1961)	1 (low prestige)	100 (high prestige)
Birth order	Birth order among siblings	1 (first born)	10 (tenth born)
Number of siblings	Number of siblings	1 (only child)	10 (ten siblings)
Liking for X	"What word best describes your liking for X?"	1 (Strongly unfavorable)	5 (Strongly favorable)
Doing well in X class (1972)	"Compared to classmates, how well are you doing in your X class?"	1 (Less well than most)	5 (Better than all)
Learning X	"How are you learning most of your X?"	1 (With my classmates)	4 (On my own, with little help)
X importance	"How important will X be for a job someday?"	1 (Not at all)	4 (Very)

Table 1: Typical items from initial assessment questionnaires, minimum, and maximum possible responses and meaning. Years next to variable names indicates which version of the item was received by the respective cohort.

Duncan (1961). Participant birth order (e.g. 1 meaning first-born and 2 meaning second-born) and number of siblings was also collected.

Participants responded to the following questionnaire items: "What word best describes your liking for school in general?", "What word best describes your liking for math class?", "Compared to your classmates, how well are you doing in your mathematics class?", "How are you learning most of your mathematics?", and "How important will mathematics be for a job someday?".

Participants also listed every grade skipped prior to identification, and this information was reduced to the simple sum of all previous grades skipped.

1976 Cohort

Table 3 lists 21 variables collected at the initial identification of the 1976 cohort. Several variables are identical to those collected in the 1972 cohort with a few exceptions.

First, both SAT-Math and SAT-Verbal subtest scores were available for this cohort. Secondly, the type of school (public, private, or church) attended by the participant at the time of collection was collected, as well as preferences, perceived class standing, and importance of several additional academic subjects. Thirdly, parental educational degrees are coded on a 1-9 scale (ranging from "less than high school" to "post-doctoral experience").

1980 Cohort

Table 4 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. All participants in the 1980 cohort were administered items assessing their preference for various academic subjects (e.g., "What word best describes your liking for math class?"), but only a subset of the participants were administered items rating the importance of subjects, class standing in subjects, or method of learning subjects. Therefore, these items were not used in the matching process due to high degree of missingness.

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 preference variables in 1972 cohort, and very few observations are missing in the two later cohorts.

Missing values are handled similarly in all cohorts. Rather than delete these cases or use mean imputation, missing values are multiply imputed using the Amelia II package for missing data (Honaker, King, & Blackwell, 2007). Multiple imputation creates m datasets, maintaining observed values and imputing m values for each missing observation (Rubin, 2004; King, Honaker, Joseph, & Scheve, 2001). Assuming that the joint distribution of the variables is multivariate normal, imputed values for each missing observation are drawn from the posterior predictive distribution. Missing

values vary across the m imputed datasets, reflecting the degree of uncertainty around each missing observation. For the current analysis, $m = 10$. Prior to imputation, variables are transformed, if necessary, to satisfy the normality assumption (as best as is possible) and are returned to their original scale after imputation. Matched samples are created in each imputed dataset.

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. This process is automated with the Zelig package in R (Imai, King, & Lau, 2007, 2009). This procedure is a more conservative approach than listwise deletion or mean imputation, as it maintains all observed data while adding uncertainty to the imputed values (Horton & Kleinman, 2007).

In the current study, variables with missing values are used to estimate individual propensity scores, and these scores are 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. All reported statistical summaries in the current analysis combine information from the 10 imputed datasets for each cohort. Simulation studies have demonstrated that this procedure of combining matching and multiple imputation greatly reduces bias compared to listwise deletion and exhibits very little bias if the data is missing completely at random (MCAR) or missing at random (MAR; Qu & Lipkovich, 2009; Crowe, Lipkovich, & Wang, 2010).

Grade Skipping

Participants responded in follow-up questionnaires, at ages 18 and 23, to items 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 high school grades skipped. Most participants 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.

The number of grades skipped after assessment was then reduced to a dichotomous variable with 0 reflecting no grades skipped and 1 reflecting one or more grades skipped. The resulting analysis then assesses directional hypotheses, comparing all grade skippers to the matched non-skippers, rather than estimating the effect of skipping exactly one grade.

Propensity Score Matching

Propensity scores, reflecting the probability of grade skipping, are estimated in each cohort based on the background variables in Tables 2, 3, and 4. In each cohort, the following procedure was used, using the 1972 cohort as an example.

First, propensity scores are estimated using a logistic regression model, predicting the grade skipping variable using all 14 background variables with only main effects. These propensity scores are then used to construct a matched control group from the total pool of control observations by finding the nearest neighbors on the

propensity score for each grade skipper. The next step is to assess the resulting balance, through inspection of the similarity of the propensity score distributions across groups (as shown in Figure 3), standardized mean differences across groups (as shown in Figure 4), and empirical quantile-quantile plots.

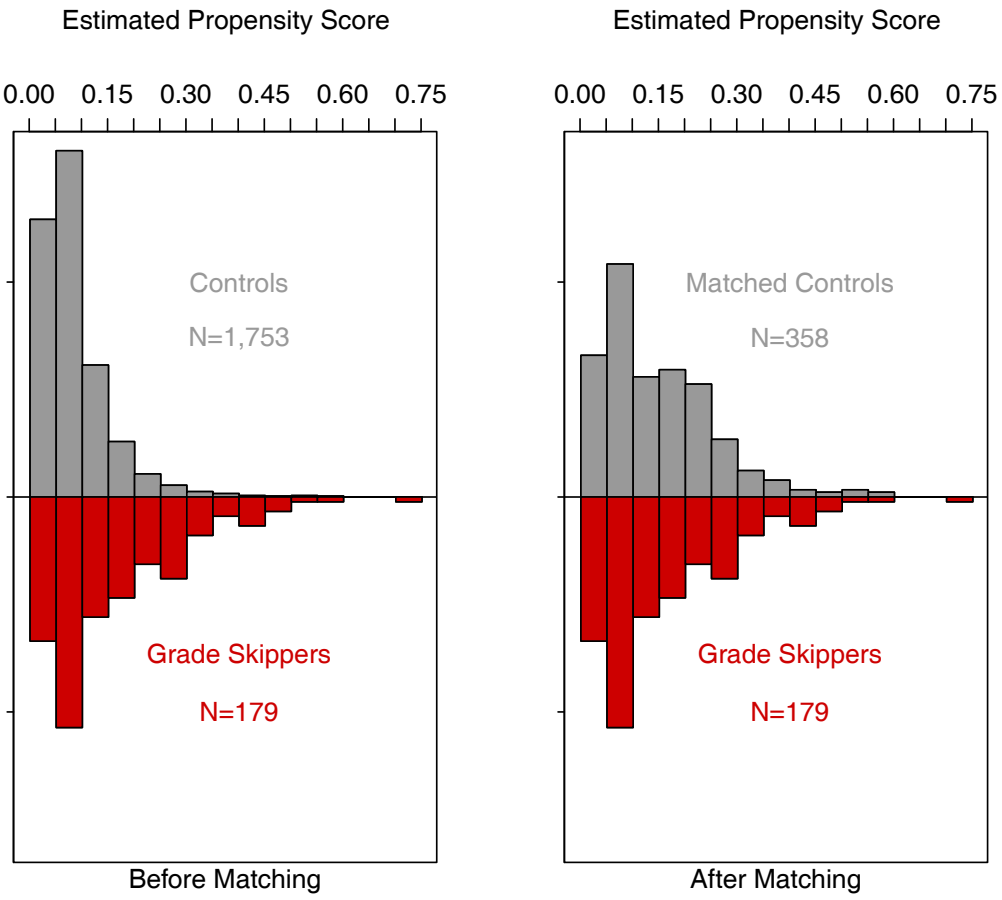

Figure 3: Example of propensity score balance between grade skipping and control groups before and after propensity score matching. Data is from the 1972 cohort. Vertical axes in the upper half of each panel are scaled to illustrate the change in distributional shape across panels.

The first propensity score model specification, using all background variables with main effects only, usually greatly improves balance, but the model can be improved through incremental changes to the model. For example, adding squared

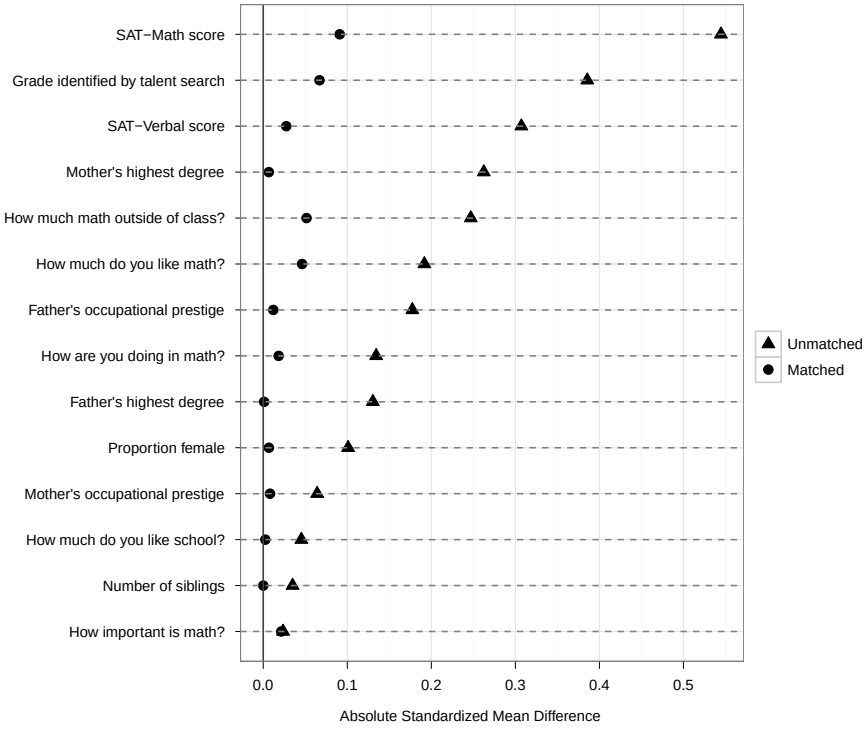

Figure 4: Absolute values of standardized differences in means on background covariates between grade skippers and control participants. Triangles indicate mean differences before propensity score matching. Circles indicate the reduced mean differences after propensity score matching. Several possible propensity score model specifications are used to reduce the imbalance in means between the two groups.

terms, interactions, or even dropping predictors from the model can result in slightly improved balance. Searching through the model space for the "best" model may initially seem counterintuitive or suspicious, as this is bad practice in more common model-fitting situations. However, the goal of finding the best propensity score model is achieving good balance in the sample, rather than parameter estimation. For this reason, models can be iteratively tweaked as long as the resulting balance continues to improve. Ultimately, balance on covariate distributions, not necessarily propensity scores, is the goal of matching.

In the current study, several propensity score models were tested until the

absolute difference in means between grade skippers and controls was approximately 0.1 standard deviations or less on all relevant background variables. An additional constraint was that grade skippers could only be matched with control participants of the same sex and with the same number of previously skipped grades. These exact matching constraints were added due to the nature of the outcomes assessed in the analysis.

Adult Outcomes

Several outcomes related to participants' educational and occupational choices and accomplishments are used to test multiple predictions of the time-saving theory.

Educational Degrees

In the current study, only post-undergraduate 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 if participants completed a dissertation or Master's thesis. Any additional information available from participants' professional website or public curriculum vita or resume was also used to determine each the educational degrees accumulated by each participants.

Degrees were coded as Master's (M.A. or M.S.), Ph.D., medical degree (M.D.

or equivalent), or law degree (J.D.). In general, a participant is coded as earning a doctorate if he/she earned a Ph.D., M.D., J.D., or a combination of these. Master's and Ph.D. 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 refers to either Master's or Ph.D.s from STEM fields.

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 of occurrence

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 each event occurrence. 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. POP automatically calculates several indices of scholarly productivity based on features such as individual's number of published articles, co-authors, and citations per article. In addition to the total number of publications and total number of citations accumulated by each participant, several common citation indices were collected for each eligible participant. From the pool of potential indices, four

indices based on their interpretability, robustness, and popularity: total accumulated citations, the h-index, the g-index, and the age-weighted citation rate.

The total number of citations of a participant is based on the total number of citations they have accrued based on all peer-reviewed STEM publications and patents on which they were listed as authors or inventors. It bluntly expressed the total impact of an individual's published work. Citation and productivity indices offer greater refinement, and while the total citation count lacks robustness (it can be easily influenced by one or two highly cited works), it has a straightforward interpretation.

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 papers 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.

Since Hirsch (2005) proposed the h-index, there has been a surge in proposed

alternative and complementary incides many other citation indices. For example, variations on the h-index include the contemporary h-index (h_c-index; Sidiropoulos, Katsaros, & Manolopoulos, 2007), the individual h-index (h_i-index; Batista, Campiteli, & Kinouchi, 2006), the normalized individual h-index ($h_{i.norm}$-index; Harzing, 2008), and the multi-authored h-index (h_m-index; Schreiber, 2008). These variations modify the original h-index by allowing the weight of older articles to decay with time (as in the h_c-index), by dividing the h-index by the average number of co-authors on every paper (as in the h_i-index and the $h_{i.norm}$-index), or by only giving partial credit for each paper to an author by dividing each paper by its number of co-authors before calculating the h-index (as in the h_m-index). These slight variations are useful when comparing individual researchers across fields and/or generations, but they are generally very similar in magnitude to the original h-index.

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 paper allows the g-index to be more sensitive to an 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, an if an author published 10 articles in the same year, five 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.

Citation counts and related indices tend to have roughly log-normal distributions with occasional severe outliers. For this reason, median values on each index are compared between grade skippers and controls for each cohort. To assess whether a statistically significant difference in the location of the distributions of each index for grade skippers and matched controls, the Wilcoxon Rank Sum test (also known as the Mann-Whitney U test) is used (Wilcoxon, 1945). The Wilcoxon test is a non-parametric alternative to a more traditional two-sample t-test and does not require any distributional assumptions, only the assumption of ordinal scaling.

CHAPTER V

RESULTS

Matching Results

By combining exact matching and propensity score matching, imbalance on many background covariates was greatly reduced. Tables 2, 3, and 4 list means on important variables measured at the initial assessment in the original control group (*unmatched controls*), the subset of the controls that were selected as matches with the high school grade skippers (*matched controls*), and the high school grade skipper group (*grade skippers*). Note that the exact matching constraints (requiring that matched control and grade skipping participants be exactly matched on both sex and number of prior grades skipped) results in perfect balance on these variables, as expected. Kernal density plots of the propensity score distributions of all controls, matched controls, and grade skippers from each cohort are shown in Figure 5.

Grade skipping participants were differentiated most notably by their SAT subtest scores. Improving balance on these variables not only decreased the differences between the grade skippers and their matched controls but between cohorts, as well. After matching, most of the standardized mean differences between the grade skippers and matched controls were smaller than .10, and all were smaller than .25, which has been suggested as the maximum allowed difference to grant the equivalency to randomization to a quasi-experimental or observational design (further adjustments

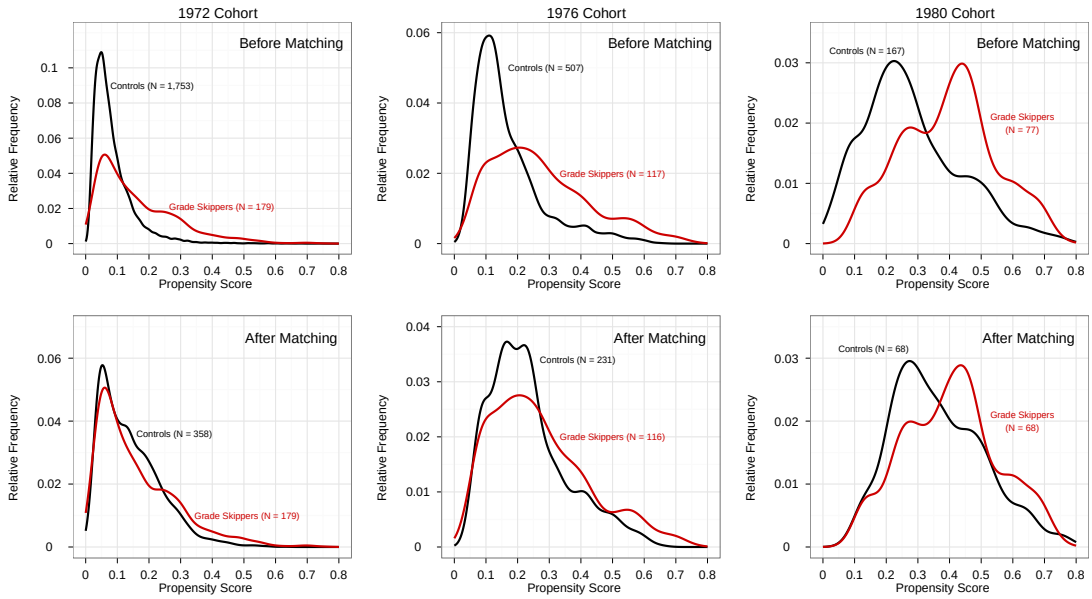

Figure 5: Kernel density plots of propensity score distributions of grade skippers and controls before and after matching. Vertical axes are scaled differently across plots.

were made using logistic regression, described below; What Works Clearinghouse, 2009; Ho et al., 2007).

Of the 2,188 participants in the 1972 cohort, 179 (102 males, 77 females) participants were identified as skipping one or more years of high school, 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. Matching two control participants with each grade skipper gave the best balance between sample size and balance.

From the 778 participants in the 1976 cohort, 116 (97 males, 19 females) participants were identified as skipping one or more years of high school, and these were matched with 231 control participants. Two control participants were matched to each grade skipper again, resulting in acceptable balance. One grade skipper could not be adequately matched with a second control participant, so the initial match for

1972 Cohort	**Unmatched Controls**	**Matched Controls**	**Grade Skippers**
N	1753	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

Table 2: Mean and proportions of 14 background variables measured at age 13 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.

1976 Cohort	**Unmatched 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

Table 3: Mean and proportions of 14 background variables measured at age 13 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.

1980 Cohort	**Unmatched 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

Table 4: Mean and proportions of 14 background variables measured at age 13 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.

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 males, 5 females) participants were identified as skipping one or more years of high school, and these were matched with 68 control participants. To maintain acceptable balance, only one control participant was matched with each grade skipper. 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 in this cohort were dropped due to lack of an acceptable match.

Comparisons of 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 5 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.

A useful summary for such comparisons is the incidence ratio, also known as the cumulative incidence ratio or risk ratio[1], interpreted here as the average increase

[1]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.

	N	Percent earning outcome			
		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

Table 5: 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.

in "risk" of reaching these outcomes due to grade skipping among the grade skippers (Cummings, 2009). Unadjusted or "crude" incidence ratios can be estimated directly from the values in Table 5 as by dividing the proportion of grade skippers earning an outcome by the proportion of matched controls earning the same outcome. Adjusted incidence ratios, which are adjusted for other observed covariates, can be estimated using a logistic regression models, either by exponentiating regression coefficients or by comparing average expected value for each subject as the grade skipping variable is changed from 0 to 1. Using the latter method, 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 6.

Each outcome has six corresponding estimated incidence ratios, each summarizing the change in risk in the grade skippers compared to the matched controls. For example, the first three incidence ratios for doctoral degrees are the adjusted

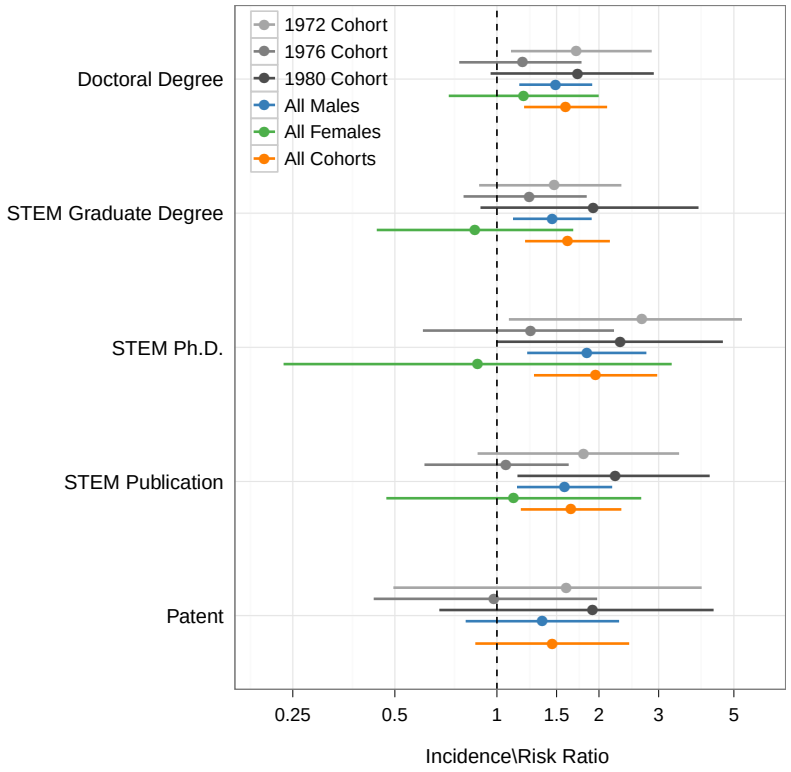

Figure 6: Estimated effect sizes, as incidence ratios, of grade skipping on five outcomes across three cohorts. Points indicate the point estimate of each incidence ratio, and horizontal lines indicate 95% confidence intervals. The vertical line at the incidence ratio of 1 indicates the point of no effect. incidence ratios for the 1972, 1976, and 1980 cohorts are adjusted for observed covariates. Pooled incidence ratios of all cohorts, all males, and all females are calculated directly by pooling across each cohort. Note: the incidence ratio for females in the patents comparison was not estimated due to lack of female participants with patents.

incidence ratios in the 1972, 1976, and 1980 cohorts, respectively. In addition, the next three incidence ratios are unadjusted incidence ratios were also calculated for all males, all females, and for all participants across the three cohorts. Assuming no confounding variables, an unadjusted pooled incidence ratio across strata of varying incidence yields an estimate of the average ratio change in risk for all grade skippers.

This property of incidence ratios, *collapsibility* (Cummings, 2009), means that incidence ratio for "all males" in Figure 6 can be interpreted in the average ratio change in risk for all male grade skippers *due to the grade skipping.*

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. However, for most individual cohorts, 95% confidence intervals around these estimates include 1, indicating that many of the estimates are not statistically significant at the using the traditional $\alpha = .05$ level.

However, the matching procedure discarded hundreds of potential control observations, trading statistical power and precision for a reduction in bias. Pooling these comparisons across cohorts reclaims some of this statistical power and summarizes the effects across each cohort. incidence ratios are significantly greater than 1 for doctorates, STEM graduate degrees (Master's and Ph.D. degrees), STEM Ph.D.s, and STEM publications but not for patents.

Limiting the pooled comparisons only to males or females 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 Ph.D.s, where female grade skippers were actually less likely than female controls to earn these outcomes. However, female grade skippers were more likely than their matched controls to earn doctorates.

		Percent earning outcome		
1972 cohort	N	M.D.	J.D.	STEM Ph.D.
Males	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
Females	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				
Males	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
Females	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				
Males	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
Females	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				
Males	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
Females	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

Table 6: 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.

Table 6 shows the patterns of percentages of different doctoral degrees across males and females, 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 6. While females were less likely to pursue STEM Ph.D.s than males, females tended to pursue medical degrees at a comparable level and law degrees to a greater extent than the males, and this differences were exaggerated among the grade skippers.

Age of Event Occurrence

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 Ph.D. 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.

		Median age of reaching outcome			
1972 Cohort	N	Doctoral graduation	STEM PhD graduation	First STEM publication	First patent
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 Controls	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

Table 7: Median ages (in years) of reaching STEM outcomes, within and across cohorts together.

Table 7 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 distribution of ages for all outcomes were positively skewed, and medians better reflect the central tendency of the distributions. 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 .9 (patents) and 1.6 (STEM Ph.D. 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.

To illustrate how these differences unfold over time, inverted Kaplan-Meier estimates of survivor functions are plotted in Figure 7 (Singer & Willett, 2003; Kaplan

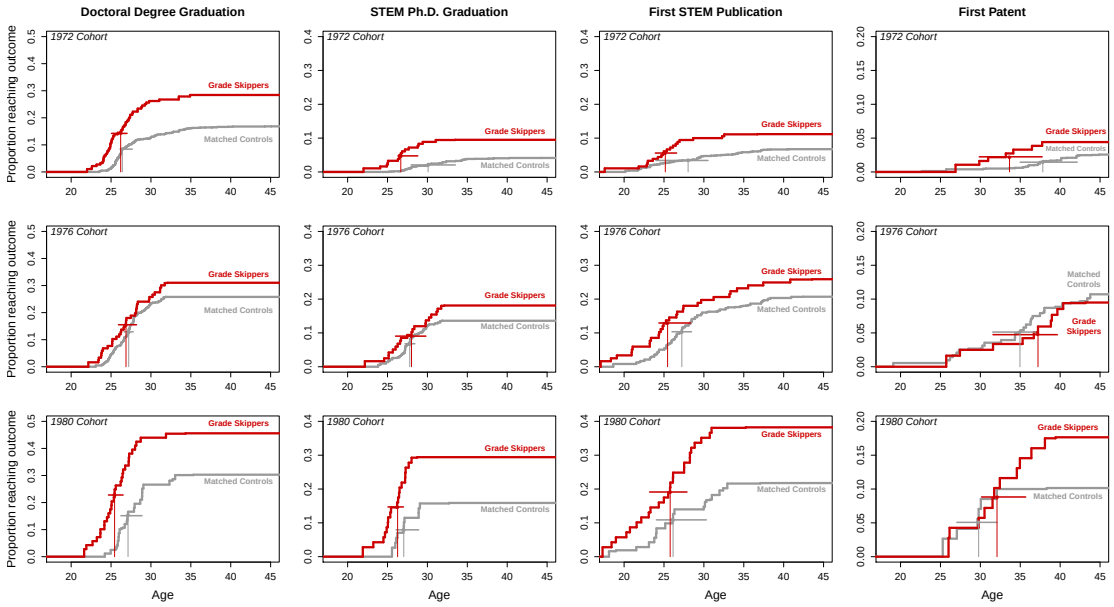

Figure 7: Inverted Kaplan-Meier estimates of survivor functions for four outcomes across three cohorts. Vertical line segments indicate the median age of event occurrence for all reaching the event in each group. Horizontal line segments indicate bootstrapped 95% confidence intervals for the medians.

& Meier, 1958). Each panel shows the cumulative proportions, in each cohort, of grade skippers and matched controls reaching each outcome as they progress from age 20 to 45.[2] Median ages within each subgroup (as listed in Table **??**) are denoted as vertical lines extending downward from each survivor function. To illustrate the variability in these medians, 95% confidence intervals are constructed around each group median using the percentile bootstrap.[3] These intervals are drawn as horizontal line segments passing through the group medians.

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

[3]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 subgroups n ages and the median of this age is calculated and recorded. This process is repeated 1000 times, resulting in 1000 medians. 95% confidence intervals are calculated using the values of the 2.5th and 97.5th percentiles of these 1000 medians.

Grade skippers tend to reach each outcome earlier, and the median ages of reaching outcomes also tend 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 Ph.D. graduation tended to have the smallest variance, with most participants finishing in their mid- to late 20s and very few graduating after age 35. The ages of STEM publications and patents were much less predictable, 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 8 to summarize the findings across cohorts. Similarly, pooled comparisons of ages show consistent age advantages of about 1 to 1.5 years for doctoral degrees, STEM Ph.D.s, and STEM publications but not patents.

Adult Productivity at Mid-career

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. Past research (e.g., Lehman, 1946, 1953; Dennis, 1956; Zuckerman, 1977; Simonton, 1988, 1997) has demonstrated a consistent relationship between the age of first accomplishment and lifelong accomplishment. In line with this research, Figure 9 shows the relationship between the age of STEM Ph.D. graduation, age of first STEM

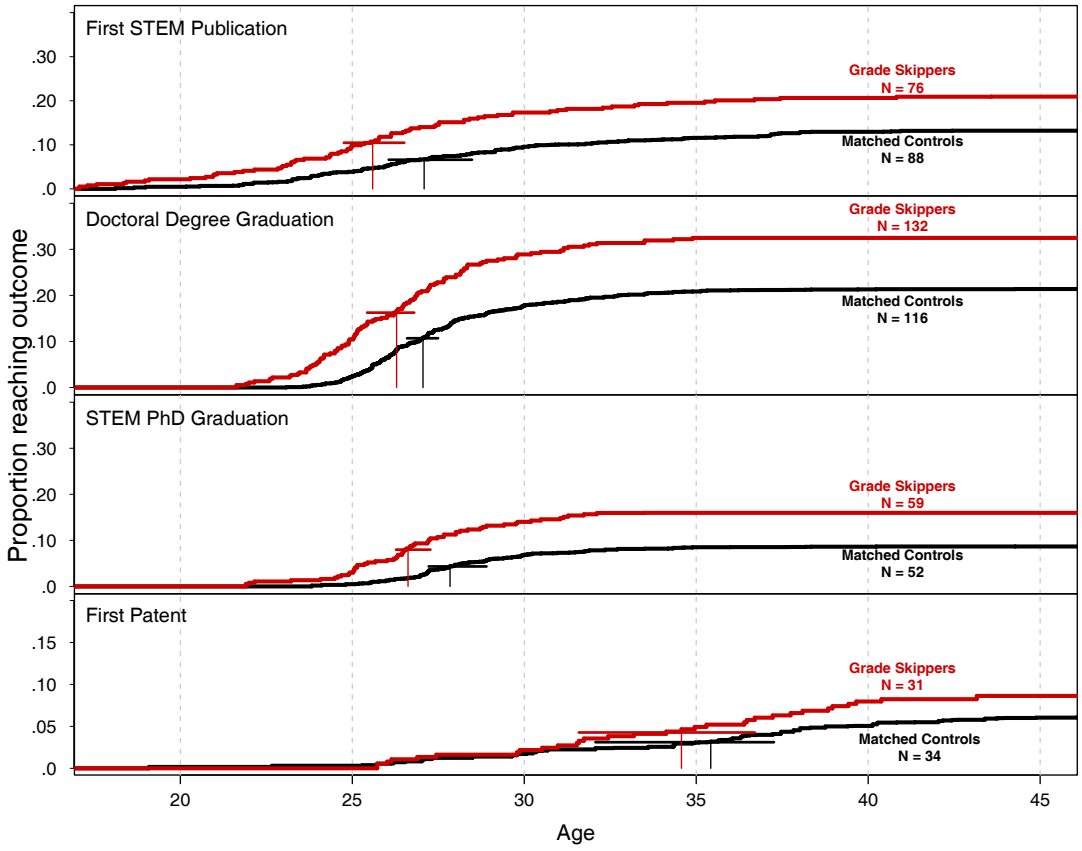

Figure 8: Inverted Kaplan-Meier estimates of survivor functions for four outcomes, pooling all three cohorts together. Vertical line segments indicate the median age of event occurrence for all reaching the event in each group. Horizontal line segments indicate bootstrapped 95% confidence intervals for the medians.

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 42, 46, and 50. Citations counts followed an approximately log-normal distribution, with many participants having citation counts in the hundreds and a few in the thousands. Due to the logarithmic scaling of the vertical

axis, small vertical distances on these plots can translate to substantial differences in raw citation counts.

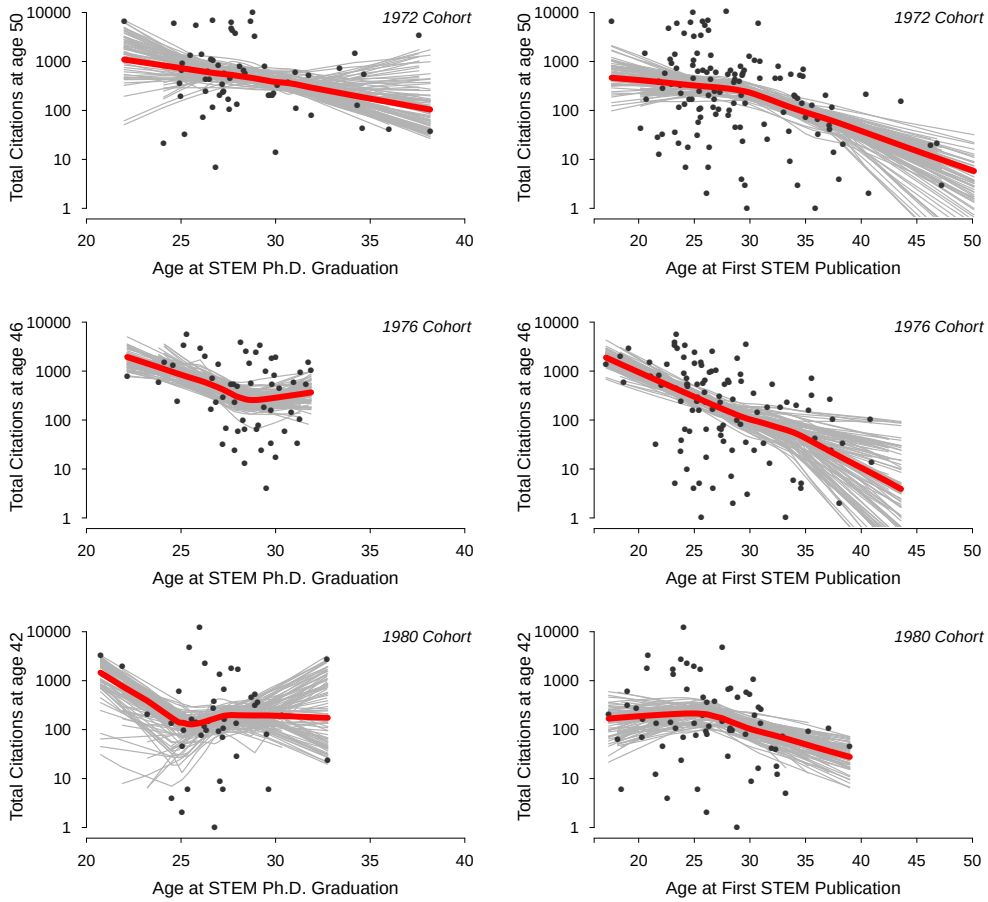

Figure 9: Scatterplots of age at STEM Ph.D. graduation and total number of citations (left) and age at first peer-reviewed STEM publication and total citations (right). Citation data was collected in 2011 when the 1972, 1976, and 1980 cohorts were 50, 46, and 42 years old, respectively. Red trend lines are fitted using a locally weighted regression (loess), and light grey lines are 100 bootstrap replications of the loess fit.

To depict trends within the clouds of points, a locally weighted regression (*loess*; Cleveland & Devlin, 1988; Cleveland, 1993) line with a wide bandwidth was drawn through each plot, shown in red. 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. The resulting fit line can allow better visualization of the general trend of the data and the presence of non-linearities. To show the stability of these fits, 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 demonstrate the robustness of the original fits (in red).

Plots in the left column show the relationships between the age of a participant's graduation from a STEM Ph.D. program and the logarithm of his/her total citation count at mid-career. The general negative trend in all three cohorts indicate that those with earlier graduations tended to have more citations in the long run. This is most dramatic between the ages of 20 and 30 and slowly levels off after age 30.

Many participants, particularly in computer science and engineering, are active researchers yet did not obtain a STEM Ph.D.. The righthand column plots the a participant's age at first publication against the logarithm of his/her total citation count. Trends are much clearer and more stable relative to the lefthand column, partly due to the increase in sample sizes (as shown in Table 5 many more participants authored a STEM publication than earned a STEM Ph.D.). The most highly cited participants tended to be those who started publishing in their early 20s.

As shown in Figures 7 and 8, the grade skipping participants tended to earn STEM Ph.D.s, author STEM publications, and earn patents earlier than their matched controls. Figure 9 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 mid-career, based on similar indices.

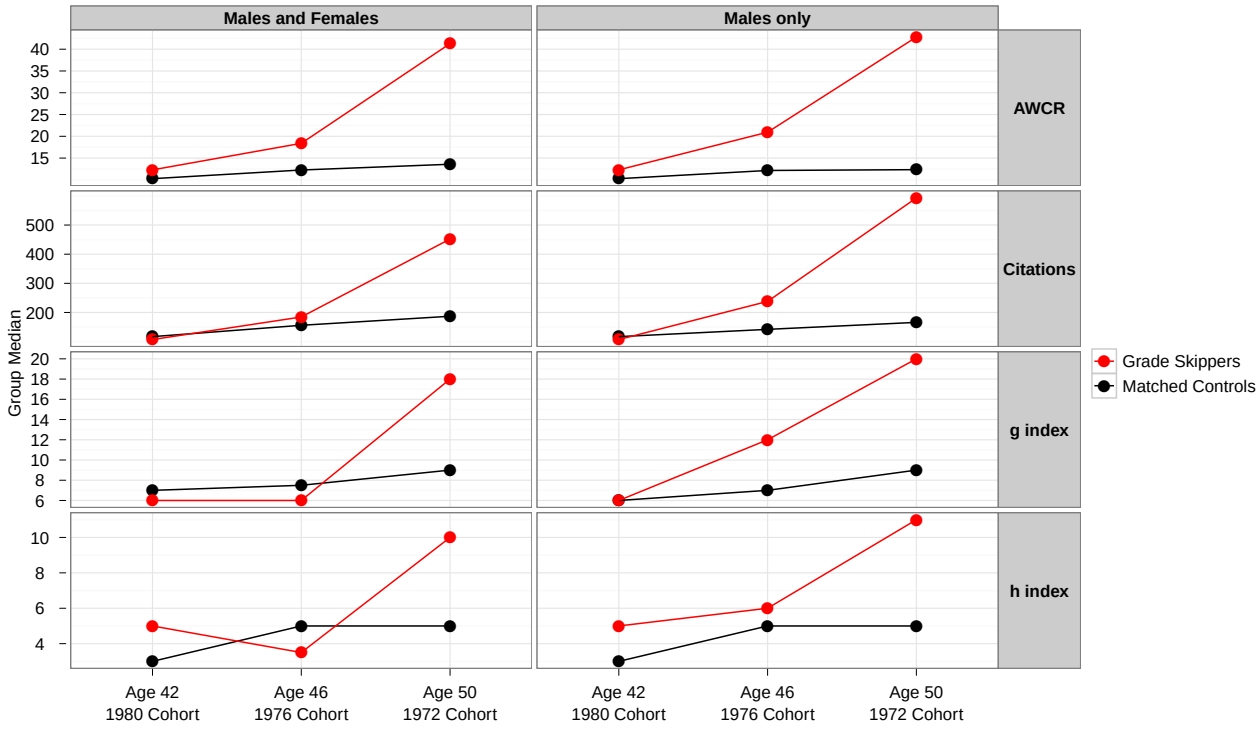

Figure 10: Productivity/citation indices from grade skippers and matched controls across three cohorts. AWCR is the Age-Weighted Citation Rate, an estimate of a participant's annual rate of citations. Citations is an participant's total number of citations accumulated from their own peer-reviewed publications 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. Ages 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.

Figure 10 plots 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 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. Total citations give a crude indication of the total influence of an individual's body of work. The h- and g-index assess productivity with a careful balance between the number of publications and citations. The AWCR provides an index of the average rate than an author is cited per year. 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 is 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 1980, 1976, and 1972 cohorts were taken when participants were at median ages of 42, 46, and 50, respectively, and these individuals are actively publishing in their respective fields. Because participants in each cohort are at different points in their careers, each cohort is plotted separately and no pooling is done across cohorts.

Inspection of Figure 10 shows a distinct advantage across all indices at age 50 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, with the matched controls slightly higher. In the 1980 cohort, taken at age 42, the opposite pattern is found, with the advantage returning to the grade skippers on most indices.

The grade skippers in the 1976 comparison contain an disproportionately high number of female authors (20.0%) compared to the 1972 (0%) and 1980 (9.7%) cohorts. Males and females 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. Males tended to publish more consistently throughout their careers. To clarify the current comparisons, the right column of Figure 10 displays only male grade skippers and their matched controls. Restricting these comparisons to males reveals a pattern of increasing advantage among grade skippers that increases from age 42 (1980 cohort) to 46 (1976 cohort) to 50 (1972 cohort). Figure 11 displays similar information in a slightly different way by only plotting *differences in group medians*.

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 Figure 10.

No differences between grade skippers and matched controls at age 42 (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 (the 1976 cohort) were also small and nonsignificant when both males and females were included ($.32 > p > .17$). Restricting the comparisons to only males increased the magnitude of these differences ($.09 > p > .05$). The largest differences between grade skippers and matched controls were observed at age 50 (the 1972 cohort). Due to the low proportion of females 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 males ($.04 > p > .01$)

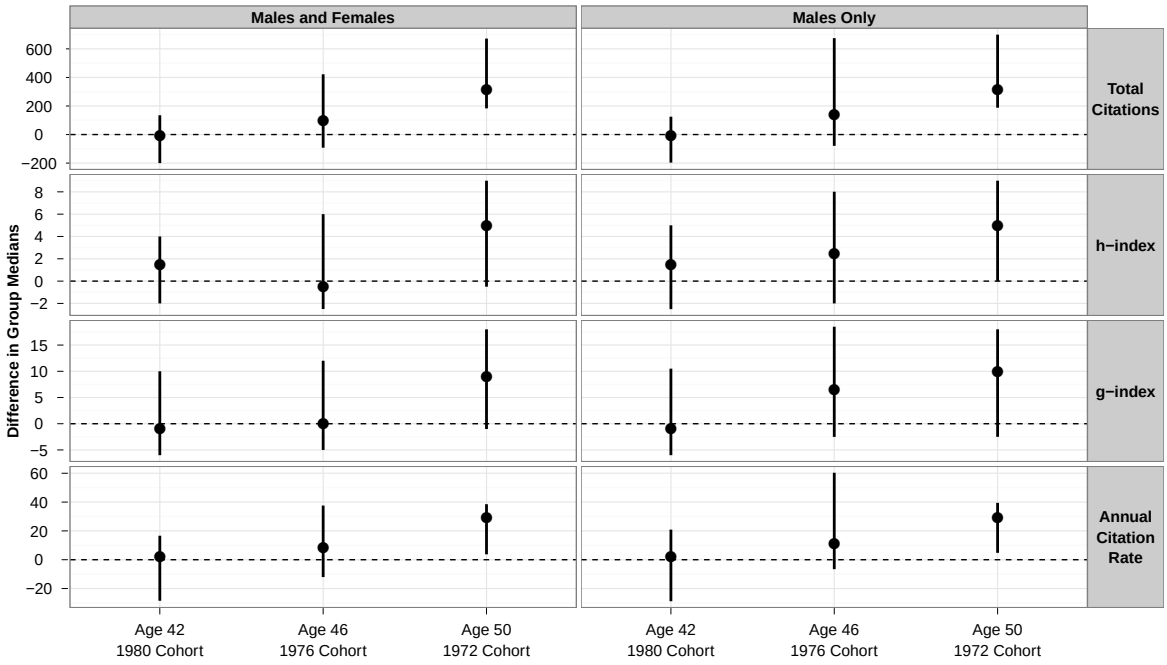

Figure 11: Median differences in productivity/citation indices, comparing grade skippers and matched controls in each cohort. Total citations refers to a participant's total number of citations accumulated from their own peer-reviewed publications and patents. The h-index and g-index are productivity 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.

CHAPTER VI

DISCUSSION

Summary

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 individuals, can have lasting effects on the productivity of those pursing STEM fields. The first phase, summarized by Figure 6, reinforces past findings in the acceleration literature (e.g., Flesher & Pressey, 1955; Pressey, 1967; Swiatek & Benbow, 1991; Wai et al., 2010; Bleske-Rechek et al., 2004). As in these previous studies of grade-based acceleration, grade skippers were more likely to pursue advanced degrees and reach important career milestones related to success in STEM careers, such as STEM publications and patents. The current study not only replicates these findings, it strengthens them by finding similar patterns of results under the much stricter methodological controls granted by the matching procedure.

Given the recent calls for increasing the STEM workforce and building STEM expertise (National Science Board, 2010a), two general findings from the first phase should be stressed. First, as shown in Table 5, both the grade skippers and their

matched controls earned highly sought achievements in STEM domains at rates several times higher than base expectation. That a relatively short test can identify a subpopulation with such potential for STEM accomplishment decades later underscores the power of early identification of mathematical talent. Educational psychologists and those in applied fields have long urged the importance of talent identification, followed by educational opportunities commesurate with their potential (Seashore, 1922; Paterson, 1957; Terman, 1954; Stanley & Benbow, 1982), and while these recommendations have already received empirical support, the findings from the current study bolster these suggestions to an even greater extent.

The second key finding from the first phase is that, while identification of mathematical talent is critical, interventions based on this identification can further optimize development of those with the most potential for STEM accomplishments. Based on 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 guess of what the grade skippers would be like had they not grade skipped. As shown in Table 5, the matched controls did not flounder without grade skipping. In fact, they earned all of the same accomplishments at very high rates, too. The matched controls are clearly at great promise for STEM achievement. What is impressive is that a relatively simple intervention, such as grade skipping, can develop this pool of talent even further.

Because grade-based acceleration acts by removing barriers to development

(such as being tethered to a lockstep, age-based educational track) rather than bringing additional content or instruction to the student, it is a good example of what Pressey (1955) described as *furtherance*, in contrast to frustration. While negative experiences, failures, and personal obstacles can hamper and frustrate development, positive experiences, successes, and opportunities have the opposite effect of spurring development on even further. Allowing mathematically talented and highly motivated students to move more freely through the existing educational track is one example of such a "furthering opportunity" (Pressey, 1955) increasing the tailwind on an already fast moving object.

The second phase of this study, 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) demonstrated that participants that 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 8 and Table **??** confirm the lasting effects of grade skipping. Grade skippers not only entered but finished their STEM graduate degrees earlier, and when criteria are broadened to include all doctorates and STEM publications, similar effects are found.

Demonstrating that grade skippers indeed reach milestones earlier than matched controls fills an existing gap between the educational acceleration literature (e.g., Stanley, 1973; Swiatek & Benbow, 1991; Pressey, 1967; Flesher & Pressey, 1955) and work on age and lifetime accomplishment (e.g., Lehman, 1946, 1953; Dennis, 1956; Zuckerman, 1977; Simonton, 1988). 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 backwards to determine the age of their first major accomplishment (Pressey, 1955). While these studies often lead to fascinating personal histories, age of accomplishment is always confounded with individual differences in abilities, motivation, and opportunities.

Figure 9 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 demonstrate that grade skipping does indeed decrease the age of first accomplishment, the story becomes clearer. The longitudinal nature of the SMPY study affords the best of both worlds, in a sense, because it is possible investigate influences on age of first accomplishment and then follow these same individuals through their early and mid-career. The critical piece of this puzzle, showing that the age of accomplishment mediates the effect on later productivity, is arguably still out of reach (Green, Ha, & Bullock, 2010; Bullock et al., 2010; Zhao, Lynch Jr, & Chen, 2010) with observational data, but the aggregate findings from all

three phases of this study constitute some of the strongest existing evidence of the effects of acceleration on adult productivity.

The final phase of the study is the first, to the author's knowledge, to study STEM accomplishments as fine-grained as citations and citation indices of STEM researchers. Past research (Park et al., 2007, 2008; Wai et al., 2010) has 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.

Narrowing the scope of the analysis to only males for greater clarity, shows a pattern consistent with this interpretation, as seen in the right column of Figure 10. Restricting the comparisons to males is necessary due to the diversity of the paths of the female participants, with many publishing early but later transitioning out of research positions into administration, teaching, or to entirely different fields. Terman (1954) reported only the outcomes of males for similar reasons.

The results from this phase, summarized in Figure 10, illustrate a pattern of increasing advantage as the cohorts increase in age. This can be interpreted in at least three ways. Due to the cross-sectional nature of these comparisons, one skeptical interpretation is that the increase in the magnitude of the differences as cohorts are

42, 46, and 50 years of age, is due to chance or cohort effects. For example, the lack of a difference between grade skippers and matched controls in the 1980 cohort may be due to the lack of effectiveness of grade skipping among the participants in that specific cohort. Or, there could be no effect of grade skipping on these outcomes in any cohort, but progressively worse matching going from the 1980 cohort to the 1972 cohort. Longitudinal data on the citation indices *themselves* is necessary to address such questions, and this data is currently not available (though it is possible to obtain).

A second interpretation is that the observed differences in citation indices reflect the different advantages among grade skippers in the age of first STEM publication. The 1972 cohort grade skippers tended to author their first publication three years earlier than the controls, while the median age difference in first publication in the 1980 cohort was only .3, or about 4 months. 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 10 are to be expected. It remains unclear why the median age differences in first STEM publication vary across cohorts as much as they do. If authoring such publications has become easier over time, then perhaps the advantages granted by grade skipping on this particular outcome have less influence over time as well.

A third interpretation, more favorable to the time-saving theory, is that the indices are relatively good "snapshots" of a similar pool of STEM researchers at ages 42, 46, and 50. If this is true, 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 example of the process is illustrated in Figure 12, 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.

However, due to the focus on STEM fields, and particularly on research in those fields, findings from this study can not fully appreciate the extent of the accomplishments from the female participants of SMPY. Results from the first phase, indicating that female grade skippers were slightly more likely to enter the fields of medicine and law rather than STEM, suggest that grade skipping may simply amplify the effects of existing preferences, including those that vary between sexes (Benbow et al., 2000; Ferriman et al., 2009; Su, Rounds, & Armstrong, 2009), providing another example of furtherance. A future study using broader criteria may shed more light on the impact of grade skipping on mathematically talented females.

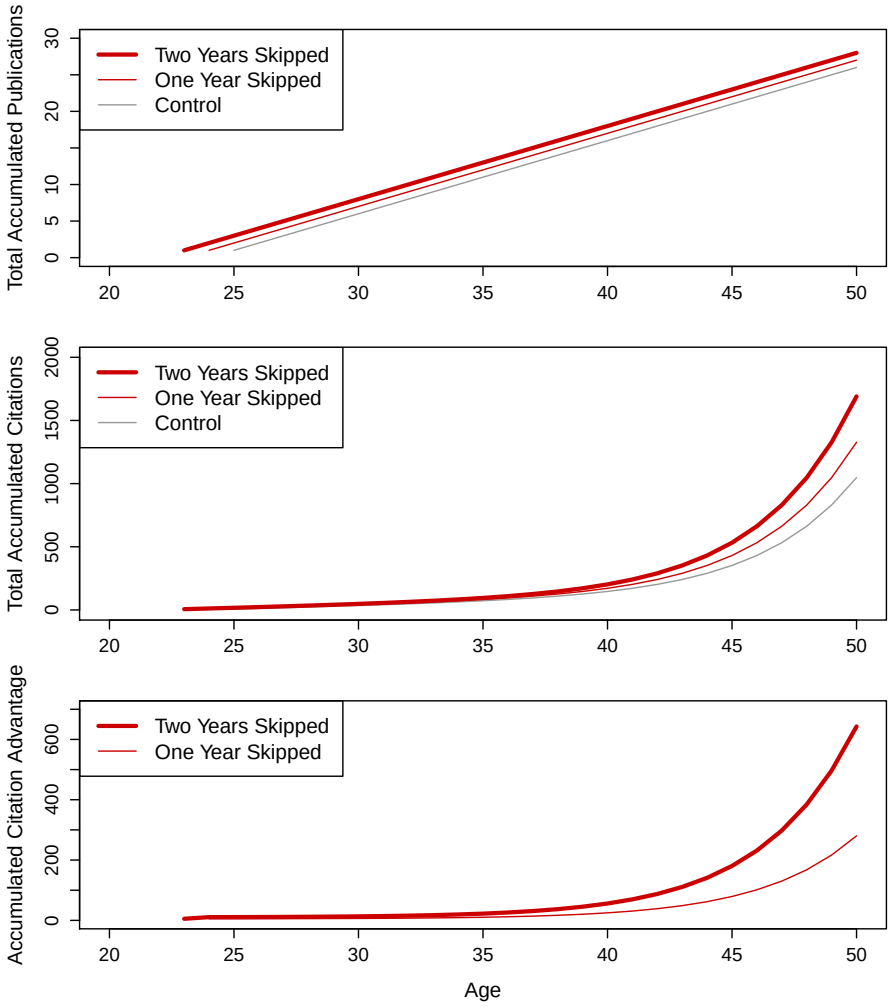

Figure 12: A hypothetical example of a small effect in initial starting points resulting in large differences in mid-career. The top panel shows the cumulative publications of of the same individual, publishing one article per year, under three possible starting ages (26, 25, and 24). The middle panel 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.

Limitations

Other limitations to this study exist. Most importantly, the matching procedure used only matches on observed variables. This leaves open the possibility that differences between the accelerates and their matched controls was, at least in part and at most completely, due to differences on unobserved variables. While the matching methodology lends considerable strength to the study of observational data, it still lacks the inferential power granted by randomization. Although findings are replicated across three cohorts, sensitivity analyses could provide an alternative approach to assessing the robustness of the findings in matched samples. Sensitivity analyses investigate the observed effect in the presence of simulated unobserved covariates and can be useful in determining how easily the effects could be explained by such an unobservable covariate under various conditions.

Additionally, matching removed hundreds of observations, decreasing power to detect effects. Precision and power were traded for a decrease in bias, and this leads to greater uncertainty in the resulting estimates. On the other hand, to paraphrase 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. Furthermore, the lack of precision in estimates was countered in this study by replicating findings across three cohorts.

Generalizability of these findings may be weakened by the over-representation of males, which was present in the original cohort samples, and further compounded by the matching procedure, focus on STEM fields, and final analysis of active researchers in these fields. Furthermore, while the longitudinal nature of the study allowed the

testing of developmental hypotheses, it also forces the analysis to consist of individuals who skipped grades in the late 1970s and early 1980s. Grade skipping is still a fairly uncommon form of acceleration compared to, say, AP courses, but selection procedures may change over time, and this could change the effects of grade skipping in an unknown way across cohorts.

Finally, the indicators of educational and occupational accomplishments heavily favored academically-oriented careers in STEM. Broader indicators at mid-career, such as career satisfaction, position in an organizational hierarchy, or income would allow for a much more comprehensive assessment of the effects of acceleration for those individuals pursuing more diverse careers outside of STEM research.

Closing

Overall, the findings from this study are supportive both of the theory concerning the mechanisms underlying acceleration's effects, as described during the peak of interest in acceleration almost 60 years ago (Pressey, 1946b; Terman, 1954; Paterson, 1957), and also of the more recent policy recommendations following earlier empirical support of acceleration (Colangelo, Assouline, & Gross, 2004; Benbow & Stanley, 1996; Stanley & Benbow, 1982). While the results fit reasonably well with the time-saving theory, they also generate additional hypotheses that may be address in future research.

Differences in accumulated productivity was assessed by comparing citation indices of grade skippers and control participants when they were approximately the

same age. Using this "snapshot" method is uninformative about the differences in individual and group growth trajectories of citations and other creative products. With the currently available bibliometric data, it is possible to create a richer longitudinal dataset tracking each individual's citation count, publication count, and corresponding indices across the span of their career. This would provide insight into a number of interesting differences in the individual trajectories, allowing comparisons of those who start careers at different ages, work in different fields, or reach varying levels of career accomplishments.

In addition, incorporating assessments of spatial ability (Wai et al., 2009), vocational interests (Ferriman et al., 2009; Su et al., 2009), would facilitate not only better matches but also the use of broader adult outcomes. Restricting the sample to the mathematically precocious yielded a sample with high potential for STEM accomplishments but at the cost of fully appreciating the diversity of accomplishments among the intellectually talented. This was especially true for the women in the sample, many of whom opted for careers outside of STEM research, in medicine, law, education, and administration. Combining more highly detailed longitudinal data, broader assessments and outcome criteria, and statistical modeling (rather than matching) may tell a much more nuanced story of development without sacrificing sample size or the diversity of individuals in the sample.

Still, the current study represents an important step in untangling interventions and individuality among the mathematically precocious and presents additional evidence that a relatively simple and low-cost form of acceleration such as grade skipping

may result in greater efficiency in education, more satisfaction among precocious students by furthering their development, and, through increased scientific productivity, benefits for society at large.

REFERENCES

American Competitiveness Initiative. (2006). *American Competitiveness Initiative: Leading the world in innovation*. Washington, D.C..

Austin, P. (2010). Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American Journal of Epidemiology*, *172*, 1092–1097.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173-1182.

Batista, P., Campiteli, M., & Kinouchi, O. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics*, *68*, 179–189.

Bellman, R. (1961). *Adaptive control processes*. Princeton University Press.

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.

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.

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.

Bullock, J., 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.

Campbell, D., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research on teaching*. London, UK: Rand McNally.

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

Cleveland, W. S., & 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. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*, 295–313.

Colangelo, N., Assouline, S. G., & Gross, M. U. M. (2004). *A Nation Deceived: How School Hold Back America's Brightest Students*. Iowa City, IA: 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, IA: University of Iowa Press.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, *112*, 558-577.

Cook, T., & Steiner, P. (2010). Case matching and the reduction of selection bias in

quasi-experiments: The relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychological Methods*, *15*, 56–68.

Correll, D. (2010, February). In Utah, a plan to cut 12th grade. *Los Angeles Times*. Available from `http://articles.latimes.com/2010/feb/15/nation/la-na-utah-school15-2010feb15`

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.

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

Dehejia, R., & Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, *94*, 1053–1062.

Dennis, W. (1956). Age and productivity among scientists. *Science*, *123*, 724–725.

Diamond, A., & Sekhon, J. S. (2006). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. (Working paper).

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

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*, 131–152.

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.

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: Houghton Mifflin.

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

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

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press Cambridge.

Gordon, R. A. (1967). Issues in the ecological study of delinquency. *American Sociological Review*, *32*, 927–944.

Gordon, R. A. (1968). Issues in multiple regression. *American Journal of Sociology*, *73*, 592–616.

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.

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

Harzing, A. W. (2011). *Publish or perish, version 3.0.4084.* Available from `www.harzing.com/pop.htm`

Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*, 165-169.

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.

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

Hobbs, N. (1958). The compleat counselor. *Personnel and Guidance Journal*, *36*, 594-602.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.

Honaker, J., King, G., & Blackwell, M. (2007). Amelia II: A program for missing data. Available from `http://gking.harvard.edu/amelia/`

Horton, N., & 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.

Huber, J. (1999). Inventive productivity and the statistics of exceedances. *Scientometrics*, *45*, 33–53.

Iacus, S., King, G., & Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, *106*, 345–361.

Iacus, S. M., King, G., & Porro, G. (in press). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*.

Imai, K., King, G., & Lau, O. (2007). logit: Logistic Regression for Dichotomous Dependent Variables. *Everyone's Statistical Software*. Available from `http://gking.harvard.edu/zelig`

Imai, K., King, G., & Lau, O. (2009). Zelig: Everyone's statistical software [Computer software manual]. Available 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.

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.

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.

Kremen, W. S., Seidman, L. J., Faraone, S. V., Pepple, J. R., Lyons, M. J., & Tsuang, M. T. (1996). The "3 Rs" and neuropsychological function in schizophrenia: An empirical test of the matching fallacy. *Neuropsychology*, *10*, 22–31.

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

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

Lewin, T. (2002, April 17). Questions for advanced placement. *The New York Times*, p. A16.

Lichten, W. (2000). Whither advanced placement? *Education Policy Analysis Archives*, *8*. (Retrieved July 1, 2010, from http://epaa.asu.edu/ojs/article/viewFile/420/543)

Lubinski, D., & Benbow, C. P. (2006). Study of Mathematically Precocious Youth after 35 years: Uncovering antecedents for the development of math-science expertise. *Perspectives in Psychological Science*, *1*, 316–345.

Lubinski, D., Benbow, C. P., Webb, R. M., & Bleske-Rechek, A. (2006). Tracking exceptional human capital over two decades. *Psychological Science*, *17*, 194-199.

Mangan, K. (2008, June). Northwestern university law school is latest to introduce 2-year degree. *The Cronicle of Higher Education*. Available from `http://chronicle.com/daily/2008/06/3488n.htm`

Meehl, P. E. (1970). Nuisance variables and the ex post facto design. *Minnesota studies in the philosophy of science*, *4*, 373–402.

Meehl, P. E. (1971). High school yearbooks: A reply to Schwarz. *Journal of Abnormal Psychology*, *77*, 143–148.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. New York, NY: Cambridge University Press.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the national mathematics advisory panel*. (Tech. Rep.). Washington, D.C.: U.S. Department of Education.

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

National Science Board. (2010b). *Science and engineering indicators 2010* (Tech. Rep.). Arlington, VA: National Science Foundation (NSB 10-01).

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.

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.

Paterson, D. G. (1957). The conservation of human talent. *American Psychologist*, *12*, 134-144.

Pressey, S. L. (1946a). Acceleration: Disgrace or challenge? *Science*, *104*, 215-219.

Pressey, S. L. (1946b). Time-saving in professional training. *American Psychologist*, *1*, 324-329.

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.

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.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1980a). Bias reduction using mahalanobis-metric matching. *Biometrics*, *36*, 293–298.

Rubin, D. B. (1980b). Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, *75*, 591–593.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers? *Journal of the American Statistical Association*, *81*, 961–962.

Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, *47*, 1213–1234.

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

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322–31.

Rubin, D. B. (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods*, *15*, 38–46.

Schreiber, M. (2008). To share the fame in a fair way, h-m modifies h for multi-authored manuscripts. *New Journal of Physics*, *10*, 040201.

Seashore, C. E. (1922). The gifted student and research. *Science*, *56*, 641-648.

Sekhon, J. S. (2007). Alternative balance metrics for bias reduction in matching methods for causal inference. Available 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.

Shadish, W. R. (2010). Campbell and Rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods*, *15*, 3–17.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association*, *103*, 1334–1343.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Macmillan.

Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, *72*, 253–280.

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

Simonton, D. K. (1997). Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review*, *104*, 66–89.

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

Southern, W. T., & Jones, E. D. (2004). Types of acceleration: Dimensions and issues. *A Nation Deceived: How schools hold back America's brightest students*, *2*, 5–6.

Splawa-Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, *5*, 465–472.

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

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

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

Steiner, P., Cook, T., & Shadish, W. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, *36*, 213–236.

Stigler, J. W., & Miller, K. F. (1993). A good match is hard to find: Comment on Mayer, Tajika, and Stanley (1991). *Journal of Educational Psychology*, *85*, 554–559.

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.

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

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.

Terman, L. M. (1954). The discovery and encouragement of exceptional talent. *American Psychologist*, *9*, 221-230.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*, 1–67.

Voglmaier, M. M., Seidman, L. J., Niznikiewicz, M. A., Dickey, C. C., Shenton, M. E., & McCarley, R. W. (2000). Verbal and nonverbal neuropsychological test performance in subjects with schizotypal personality disorder. *American Journal of Psychiatry*, *157*, 787.

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

Wai, J., Lubinski, D., Benbow, C. P., & Steiger, J. S. (2010). Achivement 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.

What Works Clearinghouse. (2009). *What Works Clearinghouse Procedures and Standards Handbook* (Vol. 20; Tech. Rep.). U.S. Department of Education, Institute of Education Sciences.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*, 80–83.

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

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