

ESSAYS IN THE ECONOMICS OF EDUCATION AND DISABILITY

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Sarah Jessica Prenovitz

May 2018

© 2018 Sarah Jessica Prenovitz

ESSAYS IN THE ECONOMICS OF EDUCATION AND DISABILITY

Sarah Jessica Prenovitz, Ph. D.

Cornell University 2018

My dissertation examines issues in determining program eligibility and participation in education and Social Security Disability Insurance program, and analyzes the effectiveness of the Andrew W. Mellon Foundations Mellon Mays Undergraduate Fellowship Program.

In my first chapter I investigate school responses to incentives created by No Child Left Behind to alter special education placement, and use these responses as instruments to estimate the effect of special education placement. I use administrative data from the universe of North Carolina Public Schools and a difference-in-difference framework in which incentives are determined by the interactions between schools' expectations about subgroup performance on the one hand and student performance and subgroup membership on the other. I find that schools use special education to target supports and services to students close to the passing threshold in reading when the school benefits from their passing. Schools also select the special education group to be higher performing when doing so benefits the school. Special education decreases attendance and has large negative effects on the math scores of some marginal students.

The second chapter explores the effect of time spent waiting for a Social Security Disability Insurance (SSDI) decision on health and well-being. Previous research has shown that SSDI in general, and wait time in particular, depresses labor market activity, but other effects are largely unexplored. I use administrative data from the Social Security Administration linked to the National Health Interview Survey, and use summary data to create instruments for wait time. A longer wait increases the number of conditions causing activity limitations and the likelihood of having current benefits at survey, and decreases the likelihood of seeking a reconsideration or having benefits terminated at survey.

The Mellon Mays Undergraduate Fellowship Program (MMUFP) aims to increase the number of underrepresented minorities entering earning PhDs, with an eye to improving their representation in academia. The third chapter evaluates whether the MMUFP increased the number of PhDs achieved by underrepresented minority students (URMs) at participating undergraduate institutions. The chapter finds no evidence that participation in the program causes a statistically significant increase in the numbers of PhDs completed by URM students, and increases greater than about one PhD per institution per cohort lie outside a 95% confidence interval of the estimates.

BIOGRAPHICAL SKETCH

Sarah Jessica Prenovitz grew up in Lexington, Massachusetts. After graduating high school in 2004, she enrolled at Oberlin College in Oberlin, Ohio and received a BA in economics (with high honors) and history in 2008. She then moved to the Washington, DC area where she worked at Mathematica Policy Research from 2008 to 2012, first as a research assistant/programmer and then as a research analyst. She moved to Ithaca, NY in 2012 to enroll in the PhD program in economics at Cornell University. She earned her MA in economics in 2015, and her PhD in 2018.

To Jenna.

ACKNOWLEDGMENTS

I thank Ronald Ehrenberg, Michael Lovenheim, and Nicolas Ziebarth for help, guidance, and occasional cheerleading throughout my time in graduate school. I also wish to thank the many individuals who offered feedback on this research, including Jody Schimmel Hyde, Gina Livermore, Maria Fitzpatrick, seminar and conference attendees, and several anonymous reviewers and referees.

The following additional acknowledgements and disclaimers apply to the second chapter. The research reported herein was performed pursuant to a grant from the U.S. Social Security Administration (SSA) funded as part of the Disability Research Consortium. The opinions and conclusions expressed are solely those of the author(s) and do not represent the opinions or policy of SSA or any agency of the Federal Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of the contents of this report. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply endorsement, recommendation or favoring by the United States Government or any agency thereof. This project is also supported by the Social Security Administration through a Disability Determination Process small grant via Policy Research, Inc. and grant #1DRC12000002-04 to the National Bureau of Economic Research.

The following additional acknowledgements and disclaimers apply to the third chapter. The Cornell Higher Education Research Institute (CHERI) receives financial support from the Andrew W. Mellon Foundation but the conclusions we express here are strictly our own. The use of NSF data does not imply NSF endorsement of the research, research methods, or conclusions contained in this paper.

I wish to thank my parents, for believing that I could do anything I set my mind to, and for their unconditional love and caring.

Finally, I thank Will for following me to Ithaca so that I could pursue a PhD, and for his support, patience, and encouragement since the day we met. And Jenna, for the joy and love she brings into my life.

Any errors or omissions are my own.

TABLE OF CONTENTS

Biographical Sketch	iii
Dedication	iv
Acknowledgements	v
Table of Contents	vi
Chapter 1. Accountability Incentives and Special Education	1
1. Introduction	2
2. Policy Background	7
2.1 Special Education	7
2.2 No Child Left Behind	9
2.3 State Accountability	11
2.4 Resulting Incentives	13
3. Prior Literature	13
4. Data	16
5. School Responses to Accountability Incentives	22
5.1 Method	22
5.2 Placement Responses to AYP Incentives	28
5.3 Heterogeneity	34
5.4 Test Taking and Selection	36
6. Effects of Special Education on Student Achievement	39
6.1 Method	39
6.2 Effects of Special Education on Achievement	41
6.3 Mechanisms	44
7. Conclusion	50
8. References	54
Chapter 2: What Happens When You Wait? Effects of Social Security Disability Insurance Wait Time on Health and Financial Well-Being	57
1. Introduction	57
2. Data and Sample	61
2.1 Data Sources	61

2.2 Sample	62
3. Method	67
4. Results.....	70
4.1 First Stage.....	70
4.2 Effects of Wait Time.....	71
5. Conclusion.....	75
6. References	77
Appendix to Chapters 1 and 2	79
Chapter 3: An Evaluation of the Mellon Mays Undergraduate Fellowship’s Effect on PhD Production at Non-UNCF Institutions.....	91
1. Introduction	91
2. Background and Program Structures.....	93
2.1 Background.....	93
2.2 The Mellon Minority/Mays Undergraduate Fellowship Program	93
3. Data and Methods	94
3.1 Data and Sample	94
3.2 Method	96
4. Results.....	96
4.1 Baseline Estimates	96
4.2 Estimates of Program Intensity.....	98
4.3 Robustness.....	99
5. Conclusion.....	99
6. Appendix A.....	100
7. Appendix B.....	101
8. References	102

CHAPTER 1

ACCOUNTABILITY INCENTIVES AND SPECIAL EDUCATION

1. Introduction

About 13 percent of students in US public schools receive some form of special education, with specific learning disabilities and speech or language impairments the most common conditions (U.S. Department of Education, National Center for Education Statistics 2016). In 2012-2013, about 19 percent of public school spending was on students who received special education, amounting to approximately \$118 billion (Federal Education Budget Project 2014, U.S. Department of Education, National Center for Education Statistics 2013). However, the effect of special education on student outcomes is not well understood. In theory, special education could either improve or hurt student achievement. Students who receive special education can benefit from extra attention and individualization, better understanding their own learning needs, or accommodations such as extra time for testing. On the other hand, they may suffer from stigma, respond to low expectations with low effort, or miss out on opportunities available to their peers in regular education (Bear, Clever, & Proctor, 1991, Lackaye & Margalit 2006).

Public schools are legally required to provide a free and appropriate public education to students with disabilities. Decisions about whether a student is disabled and what services are needed are made through a complicated interaction of many stakeholders – teachers, administrators, school-based specialists, doctors, parents, lawyers, and sometimes the students themselves. Although students' eligibility for special education is based entirely on their impairments and educational needs, prior work has shown that the size and composition of the special education population responds to school incentives. These incentives have included those created by funding formulae and by early accountability

policies (Cullen 2003, Kwak, 2010, Cullen and Reback 2006, Figlio and Getzler 2006, Cohen 2007, Bokhari and Schneider 2011, Chakrabarti 2013, Jacob 2005, Mahitivanichcha and Parrish 2005, Winters and Greene 2011, Hanusheck and Raymond 2005, Morrill, 2016).

Under early accountability programs students enrolled in special education were not included in the accountability population and were in many cases exempt from testing. This presented schools facing accountability pressure with an incentive to encourage low-performing students to enter special education. No Child Left Behind, a federal law enacted in 2002, required all states to introduce accountability programs that held schools responsible for the performance of all students as well as student subgroups defined based on demographics. One of these subgroups was the group of students with disabilities (SWD). A school made adequate yearly progress (AYP) only if all of its subgroups, including students with disabilities, met targets for participation, proficiency, and either attendance or graduation rate.¹

These new accountability programs presented schools with at least two incentives to alter the assignment of students to special education. First, schools could use special education placement to target services, individualization, and testing accommodations to “bubble” students expected to be near the passing threshold, particularly those in subgroups that were expected to fail to make AYP. Second, schools may have tried to select their special education population to be higher-performing in order to make it more likely that the SWD

¹ Beginning in the 2011-2012 school year states were granted waivers from major requirements of NCLB, and it was replaced in 2015 by the Every Student Succeeds Act (ESSA). However, the accountability systems currently in place maintain the features of NCLB that underpinned its incentives for schools to alter the special education population. They judge performance at least in part based on the percentage of students who pass a given cut-score, include nearly every student in testing and accountability, and count the special education population as a separate subgroup that must meet standards.

subgroup made AYP. In this scenario, schools that had previously failed to make AYP for the SWD group would be more likely to assign a student to special education if they expected that student to achieve a passing rather than a failing score, especially if the SWD group was close to the AYP threshold. It is unknown to what extent schools respond to accountability policies put into place since the enactment of No Child Left Behind (NCLB), which were crafted in part to eliminate the incentives in early accountability programs.

I use student-level administrative data from all public school students in grades 4-8 in North Carolina from 2007-2011 to examine how schools responded to accountability incentives under NCLB to classify particular students into special education. I do so by looking for evidence that schools use special education to target services to “bubble” students near the passing threshold when the school’s AYP performance could be improved by their passing, or that schools select the special education population to be relatively high-performing when the school expects that the SWD subgroup will fail to make AYP. I use a difference-in-difference framework in which incentives are determined by the interactions between schools’ expectations about subgroup performance and students’ performance and subgroup membership. My analysis is focused on students with relatively malleable diagnoses, such as learning disabilities, speech and language impairments, and ADHD. I find evidence that schools used special education to target supports and services to bubble students in reading when the school would benefit from that student passing, with students who previously scored in the achievement level just below passing about 1 percentage point more likely to be in special education.² Students just below the passing threshold in math are less likely to be in special education, while those just above the passing threshold in math are

² I consider schools able to benefit from a student passing if that student was a member of a subgroup that had previously failed to reach the AYP passing threshold.

unaffected. This pattern is probably shaped by the fact that math scores are generally easier to alter through instruction than are reading scores, as well as the fact that all schools had incentives to increase their overall passing rate. Also, schools face funding incentives to limit the overall size of the special education population, so may discourage some students from receiving special education in order to make room for others.

Responses to the second incentive are clearer. Compared with students who had previously failed, schools were relatively more likely to place previously passing students in special education when they were trying to improve the likelihood that the students with disabilities group achieved AYP, particularly in reading. A student who had previously passed their reading test would be about 2 percentage points more likely to be in special education, relative to a prior-failing student, if their school faced the highest incentive to select the SWD group based on reading versus the lowest incentive. This finding is fairly robust across specifications, but the pattern in math is weaker and not consistently significant. Responses to the second incentive appear to be driven primarily by fewer prior-failing students being in special education when their school faces accountability pressure, rather than more prior-passing students.

I then use differences in the second set of incentives across schools and students as a source of plausibly-exogenous variation in the likelihood a student was assigned to special education to estimate the effect of special education on test scores in an instrumental variables framework. Special education decreased math scores by more than a standard deviation for students whose placements were driven by accountability incentives in math. Estimates for the reading scores for these students are similar in magnitude but not significantly different from zero, while those for students whose placements were altered by incentives to improve the performance of the SWD subgroup in reading are not statistically

significant. I investigate several mechanisms and fail to find evidence of changes to grade retention or school switching, but do find changes in student effort as measured by attendance. This suggests that the lower engagement among special education students that has been documented previously is at least partially caused by special education placement, and has consequences for achievement.

This paper makes two main contributions to the literature. First, I extend our knowledge of how schools alter the assignment of students to special education in response to incentives. I do so by offering the first estimates of school responses to the special education incentives presented by NCLB, which are similar to those in current accountability policies. Special education placement should depend only on a student's impairments and needs, so any response to these incentives is important to understand. Prior work on this topic has focused on pre-NCLB policies in which schools had a straightforward incentive to place low-performing students into special education (Jacob 2005, Cullen and Reback 2006, Figlio and Getzler 2006, Cohen 2007). No Child Left Behind and current accountability policies were designed in part to eliminate this incentive, so the opportunities for schools to strategically change special education placement are more nuanced and targeted different groups of students.

Second, my estimates of the effect of special education placement on marginal students adds to a very sparse literature on the subject that has relied on strong assumptions for identification. Hanushek, Kain, and Rivkin (2002) examine students who move in and out of special education programs using student fixed effects and find that special education improves math scores. Their identification rests on two assumptions that I can relax: that any omitted variables that are correlated with both achievement and placement are static over

time, and that changes in achievement do not cause changes in placement.³ The paper most similar to this analysis is an unpublished working paper by Cohen (2007), who constructed instruments for special education placement based on Chicago's accountability program in the 1990s.⁴ Her results are too noisy to draw conclusions about what effects, if any, placement has on student achievement. Cohen's analysis also rests on the assumption that schools that faced pressure to increase the percentage of students who performed at grade level did not undertake other measures, aside from encouraging special education placement, that would have improved the performance of low-achieving students.⁵

My findings suggest that near-universal testing requirements and an emphasis on the performance of malleable subgroups eliminates one set of incentives – to put low-achieving students in special education – but creates another – to target resources to bubble students and select the SWD group to be high-performing. This is a useful lesson for accountability design, particularly as these features have continued beyond the end of NCLB. I also find evidence that special education can harm achievement for some of the marginal students whose placement is altered in response to accountability pressure. My results suggest that it is important to ensure that special education is appropriately targeted. More research is needed to understand the mechanisms that underlie heterogeneity in the effect of being placed in special education.

³ They are also missing data on the substantial portion of special education students who did not take standardized tests during this period. This could potentially induce selection bias, in either direction, depending on which students did not take tests.

⁴ Chicago's accountability policy placed elementary schools on probation if less than 15 percent of students performed at grade level, defined as scoring least at the 50th percentile on the Iowa Test of Basic Skills (ITBS).

⁵ A similar assumption would be needed in order to use the first incentive I consider – that to target resources to “bubble” students when the school would benefit from their passing – as an instrument for special education placement. Because it seems unlikely that this exclusion restriction would hold, my IV analyses use only the second incentive – that to select the SWD group to be higher performing.

The rest of the paper proceeds as follows. I present background information on special education, accountability, and other relevant policies in Section 2. In Section 3 I summarize prior research. Section 4 describes the data and sample. Next, in Section 5, I discuss the method used to estimate school responses to AYP incentives and the results of this analysis presented. In Section 6, I present the method used to estimate the effect of special education on achievement, the results of the analysis, and an investigation into potential mechanisms. Section 7 concludes.

2. Policy Background

In this section I discuss several policies and institutions. These include special education, No Child Left Behind, and North Carolina's state accountability policy. I then describe the resulting incentives to alter the special education population.

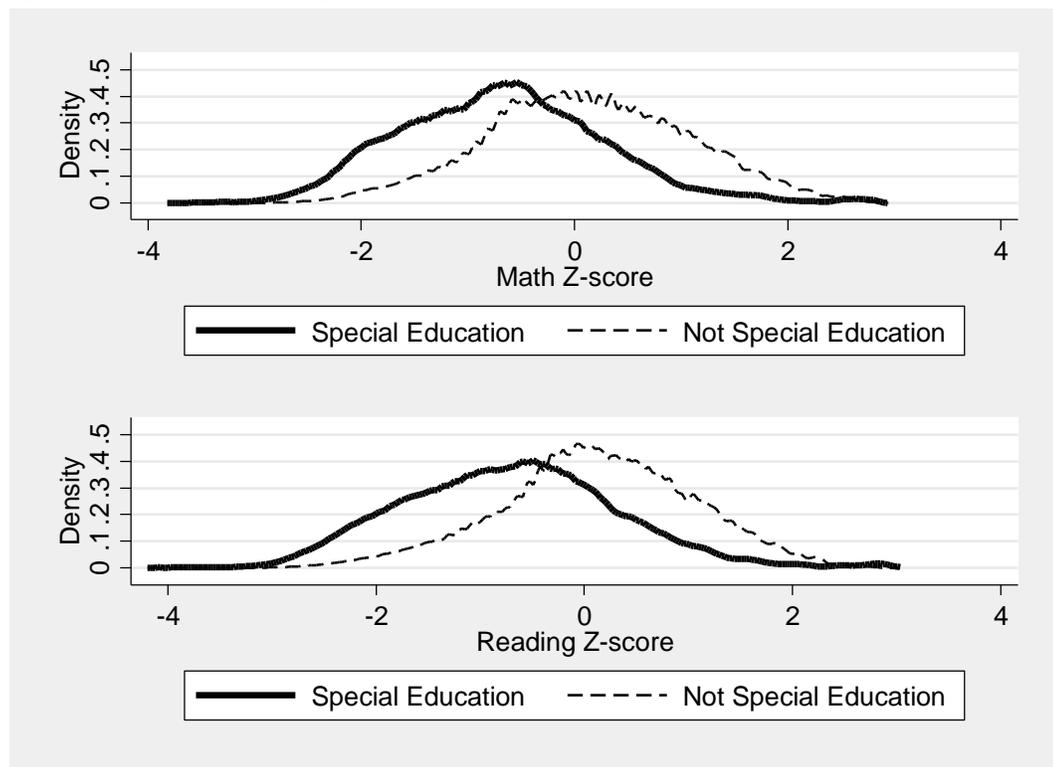
2.1 Special Education

Under the Individuals with Disabilities Education Act (IDEA) public schools must provide a free and appropriate public education (FAPE), delivered in the least restrictive setting possible, to students who are diagnosed with one of 13 categories of disability (e.g. specific learning disability, autism, visual impairment) that impedes their ability to learn or participate in other age-appropriate activities.⁶ The nature and extent of services vary widely depending on the student's needs – one student might receive weekly speech therapy, while another attends regular classes accompanied by a 1-on-1 aide, and a third student spends his time in a separate program. Special education students are lower-performing on average than their regular education peers, but there is substantial overlap between the two groups

⁶ Some students who do not have one of the 13 impairments listed under the IDEA receive accommodations under Section 504 of the Rehabilitation Act of 1973, which covers a broader set of conditions with a looser legal framework. I focus on students who are covered by the IDEA.

in terms of achievement, as shown in Figure 1. In the first panel, the distribution of math scores for students in special education is drawn with a bold solid line, and that for students not in special education with a thin dashed line. The x-axis is in standard deviation units relative to the average scores across all students.⁷ The second panel of Figure 1 displays a similar pattern using reading scores.

Figure 1. Math and Reading Scores by Special Education Status



Notes: Figure shows the distribution of math and reading scores, standardized to have mean zero and standard deviation 1 for each grade-year-subject grouping, graphed separately by special education status.

Before a student is placed into special education someone – often a parent or teacher – notices that the student is struggling and requests a disability assessment. The school then conducts a disability evaluation, which assesses the students’ abilities and needs across

⁷ These scores reflect the normalization described in detail in section 4. Dropping scores from alternate tests, rather than normalizing them to be comparable with regular tests, results in very similar patterns.

multiple dimensions. A group of stakeholders meets to establish an individualized education program (IEP), which details the services and supports the student will receive, as well as the setting in which they will be provided. These meetings include parents, teachers, administrators, specialists, and sometimes lawyers or the students themselves. Diagnoses are reviewed at least every 3 years, and IEPs every year.

Most states provide funding for special education to local education agencies (LEAs, essentially school districts) either based on the number of students enrolled or the number of special education students. North Carolina is alone in providing special education funding calculated as a set dollar amount multiplied by either the number of students with IEPs or 12.5 percent of LEA membership, whichever is smaller (Morrill, 2016). This unusual funding structure makes North Carolina a uniquely useful setting, as it can offer insights into school behavior under both common funding mechanisms. Schools in states that provide funding based on the total number of students face funding incentives similar to those faced by schools in North Carolina LEAs with more than 12.5 percent of their students in special education. Schools in states that provide funding based on the number of special education students are in situations more similar to those of schools in North Carolina LEAs below the 12.5 percent funding cap.

2.2 No Child Left Behind

No Child Left Behind, a federal law enacted in 2002, required states to establish testing programs and evaluate schools based on students' math and reading performance. States were given some leeway in determining implementation details. This paper focuses on NCLB as implemented in North Carolina, primarily from 2006-2007 – 2010-2011, so I concentrate here on characterizing that version of the policy. This time restriction allows me

to work with a relatively consistent set of policies and offers advantages in data availability, described in Section 4.

Under NCLB, each school receiving Title I funding was accountable for the performance and participation of the overall population of students as well as several subgroups defined by race/ethnicity, income, and disability status. About half of US public schools receive Title I funding, which is available to schools and districts with relatively high poverty as measured by participation in the Federal School Lunch Program. For a school to achieve Adequate Yearly Progress (AYP), at least 95% of the students in each group were required to contribute scores, the percentage of students demonstrating proficiency needed to meet target levels, and the school had to show progress on the other academic indicator (OAI): attendance and/or graduation rate. If a subgroup had fewer than 40 students it was not considered, with the exception of the full student sample.

Schools faced no consequences in their first year of AYP failure, but those that failed to make AYP in subsequent years could face sanctions. These included being forced to allow their students to choose a different school, to provide extra services, or to undergo major restructuring, depending on the number of consecutive years AYP had not been achieved. Adequate Yearly Progress was determined separately for both reading and math, such that a school could be in year 1 of AYP failure for one subject and year 3 for the other. Consequences were based on the higher of these two numbers.

There were several conditions under which a school that had not achieved all the AYP requirements would be treated as though it had done so. Schools that achieved the participation requirement and made progress in the other two measures could receive “safe harbor” and avoid sanctions. Consequences also were withheld if the proficiency threshold was within a 95% confidence interval of the actual level of proficiency or if the proficiency of

the population eligible for or receiving Title I targeted assistance met the threshold. Students who had exited LEP status or special education in the previous two years could also be included in proficiency counts. All of these details resulted in variation across schools, years, and subjects in whether an AYP failure would result in sanctions, in addition to variation across subgroups in whether the subgroup could be expected to make AYP. I exploit this variation, in addition to differences across students in subgroup membership and expected performance, in order to identify school reactions to incentives and the effect of special education on student achievement.

2.3 State Accountability

In North Carolina, NCLB operated in tandem with the state's own "ABC" program, which was first implemented in 1996 and continued with slight alterations over the period studied. Under "the ABCs," schools were labelled with various positive and negative terms based on individual student growth and the percentage passing. For example, in 2006-2007, schools that met AYP, met expected growth, and had at least 90% proficiency across grades and subjects were labelled as Honor Schools of Excellence (NCDPI 2007).⁸ Among schools with at least expected growth, those with at least 90% proficiency were labelled as Schools of Excellence, those with 80-89% proficiency as Schools of Distinction, those with 60-79% proficiency as Schools of Progress, and those with less than 60% proficiency as Priority Schools. For those schools not making expected growth, those with at least 60% proficiency were labelled No Recognition Schools and those with 50-59% proficiency were Priority Schools. Those with less than 50% proficiency that did not make expected growth were labelled as Low-Performing Schools, and were provided with state assistance such as

⁸ A school met expected growth if, on average, students at least maintained their achievement level from the prior year.

additional professional development. Schools that achieved high or expected growth could receive teacher bonuses.

While many of the principles underlying NCLB and the ABCs are similar, the precise inputs to the accountability formulae, the relative emphasis on growth vs. proficiency, and the cut-points are different across the two regimes. The ABCs also does not hold schools accountable for the performance of subgroups defined based on demographics or special education status. Thus, incentives created by the ABCs should be uniform across students with similar performance, regardless of their subgroup membership or the AYP performance of their school. However, it is possible that school responses to AYP incentives would have been different if they were not also trying to react to the state program, as the ABCs incentivizes increases in the percent proficient regardless of subgroup membership or the school's past performance.

Two additional features of the North Carolina context deserve mention. First, schools that did not receive Title I funding were evaluated under the NCLB standards. Their performance, overall and for subgroups, was announced, but there were no consequences directly tied to whether or not these schools made AYP. These schools were evaluated according to NCLB standards in all states, but states varied in how they used this information. I focus on the experiences of Title I schools, which faced stronger incentives that were consistent across states. Second, North Carolina qualified for an early waiver to the standard NCLB framework beginning in 2005-2006. This allowed students to count towards the school's proficiency rate if they either performed above the cut score or were exhibiting growth that suggested they would reach proficiency within four years of their initial test. This was uncommon when introduced by became more common elsewhere by the end of NCLB. I incorporate this detail into one measure of school expectations of student performance and

find that it does not alter my results substantially. While these two policy details are important for understanding my analysis, they do not seriously limit its generalizability.

2.4 Resulting Incentives

Prior to NCLB, accountability programs allowed schools to exclude special education students from their accountability populations and often from testing. Thus, a school could appear to have improved its performance by steering low-performing students into special education.

In contrast, there is no such option under NCLB and subsequent policies. Instead, schools could encourage special education for “bubble” students who are expected to be close to the passing threshold as a way of targeting services and supports to them. This could improve the school’s AYP performance if it expected to otherwise fail to make AYP for at least one subgroup of which the student was a member. Schools that expected to fail to achieve AYP for the SWD subgroup had an incentive to try to improve its performance. One way of doing this would be to change the group’s composition by encouraging special education for students who were expected to pass and/or discouraging those who were expected to fail. This strategy would be most useful to schools that were close to the AYP threshold, so they could change their rating by moving a relatively small number of students.

3. Prior Literature

My analysis is related to two strands of prior literature. The first has analyzed how schools respond to incentives. The part of this literature that is most closely related to this paper has addressed accountability policy incentives to alter special education placement. All previous work has focused on those in force before NCLB. These policies, examined by Jacob (2005), Hanushek and Raymond (2005), Cullen and Reback (2006), Figlio and Getzler (2006), Cohen (2007) and Bokhari and Schneider (2011), held schools accountable for student

performance but, importantly, allowed special education students to be excluded from the accountability population and often from testing. Thus, a school could increase its chance of passing by placing low-performing students into special education, and schools facing accountability pressure did just that. When programs monitored the performance of subgroups, members of those groups at risk of not meeting benchmarks were more likely than other students to enter special education, especially for students whose exclusion from accountability improved the school's performance (Cullen and Reback 2006). Using state-level variation in the roll-out of accountability policies pre-NCLB, Hanushek and Raymond (2005) found no evidence that these policies increased special education rolls, but Bokhari and Schneider (2011) found that accountability systems that provided rewards for good performance increased the number of ADHD diagnosis in the public school population, as well as the use of medication. None of this research has considered responses to more recent accountability policies, which were crafted in part to eliminate these incentives.

More broadly, a rich literature has explored how schools responded to NCLB. These responses include focusing on tested grades and subjects, focusing on the needs of "bubble" students whose passing status might change as a result, altering the testing pool through discipline, and even altering the content of school lunches on testing days (e.g. Figlio, 2006; Figlio and Winicki, 2008; Griffith and Scharmann, 2008; Krieg, 2008; Reback, 2008; Byrd-Blake et al., 2010; Dee and Jacob, 2010; Ladd and Lauren, 2010; Neal and Schanzenbach, 2010). I add to this literature by considering responses along a different margin, that of special education placements.

Prior work also has shown that schools respond to financial incentives to alter the size and composition of their special education population. These incentives can come in the form of state funding formulas (Cullen, 2003; Kwak, 2010; Mahitivanichcha and Parrish,

2005; Morrill 2016) or voucher programs open only to students with disabilities (Winters and Greene, 2011; Chakrabarti, 2013). I contribute to this research by presenting evidence on how financial and accountability incentives interact.

The second strand of literature addresses how students are affected by special education assignment. Only two previous papers have applied rigorous research designs to the question directly.⁹ Hanushek, Kain, and Rivkin (2002) used student fixed effects in a panel dataset from Texas in the 1990s and found small but significant gains in test scores in years in which students were in special education. Their identification strategy requires two assumptions that I am able to relax. First, they assume that any omitted variables that are correlated with both special education status and achievement are static, which would not be true if students' impairments change over time. Second, they must assume that changes in achievement do not cause changes in special education placement, at least after controlling for observable factors. This would be of particular concern if students are more likely to be placed in special education when struggling and to leave special education when performing well, so that regression to the mean would appear as a positive effect of special education.

In an unpublished working paper, Cohen (2007) used the accountability policy implemented by Chicago Public Schools in 1996 to construct instruments for special education placement. She found evidence that schools responded to incentives to place low-achieving students into special education but was not able to detect effects on attendance,

⁹ While not a direct analysis of the effect of special education on achievement, Setren (2016) demonstrates that special education students who win charter lotteries experience gains similar to those of their classmates who were not previously in special education, despite charters' practice of removing special education classifications at a high rate. Multivariate regressions suggest the removal of special education classifications is either not harmful or improves scores.

graduation, or GPA. Her analysis also rests on an assumption that schools trying to improve the percentage of students scoring at grade level would not do anything, aside from altering special education placement, that would affect the outcomes of low-achieving students.¹⁰ While this may be true for the very lowest achieving, who saw the greatest increase in special education placement, it is less likely for students only slightly below the average, whose probability of being in special education also increased.

While previous work has addressed many ways that schools responded to a variety of NCLB incentives, and school responses to the special education incentives in accountability policies that existed prior to NCLB, this paper is the first to consider school responses to the incentives in NCLB to alter the special education population. In doing so I am able to evaluate to what extent recent accountability policies have solved the problems identified in the earlier literature on incentives to alter the special education population. I also provide an estimate of the effect of special education on student achievement, contributing to a small literature based on strong assumptions that I relax. My estimates are local to students whose placements can be altered by schools. While this limits their generalizability it also means that they are relevant to the very group of students for whom it is most important to know the effect of special education.

4. Data

I use restricted-access student-level information from the North Carolina Education Research Data Center (NCERDC) and public use school-level information from the North Carolina Department of Education and the Common Core of Data. Student-level files provide

¹⁰ I would need to make a similar assumption in order to use the incentives to target services to bubble as instruments for special education. For this reason I do not do so, and instead only use the incentives to select the special education population as instruments.

year-by-year information on tests taken, standardized test scores, testing accommodations, disability classifications, and demographics. These files are linkable across years to create a panel that includes the universe of North Carolina public school students who were in tested grades during the years I examine. I also draw information on which schools a student attended in which grades and year from the student-level files. To the student data I add information on school characteristics and the number of years of AYP failure each school had in math and reading for all students and subgroups.

Based on their scores, students are assigned to one of four achievement levels numbered 1-4, defined by whether students have mastered grade-level content sufficiently to be prepared for the next grade.¹¹ Students in achievement levels 3 and 4 are proficient, while those in levels 1 and 2 are not. Under North Carolina's growth model, students who are in levels 1 or 2 but are on track to be proficient within 4 years of initial testing can be considered proficient for purposes of determining AYP. In my main specification, I assume schools expect students to perform about as well in the current year as in the past year, so this growth component is not relevant. However, I take the growth model into account when considering whether schools treated students who would need small gains to achieve proficiency differently from those who would need larger gains and find similar results. Thus, this modeling assumption does not drive my results.

North Carolina offered a series of alternate tests to students for whom the standard test was inappropriate, including special education students whose IEPs specified that they would take these tests. The alternate tests mean I have access to information on almost all students in the tested grades, but scores must be standardized to allow for comparisons

¹¹ North Carolina has since switched to a 5-category classification, but used this 4-category system during the period I consider.

across tests. To do so, I first assume that students who scored at a given achievement level cutoff have the same achievement – that is, students who just received passing scores for a given grade, year, and subject had the same achievement, regardless of test taken.¹² Then I assume the distance between achievement levels has the same meaning for all tests for a given grade, year, and subject – that is, students who scored halfway between the level one and level two cut points have the same achievement, regardless of test taken. Finally, I form z-scores by subtracting the mean and dividing by the standard deviation for each grade, year, subject combination. This produces a set of scores with mean zero and standard deviation one for each grade, year, and subject.

Special education serves students with a variety of disabilities. The distribution of diagnoses for special education students in North Carolina in grades 4-8 during my sample period is reported in Table 1. Schools are unlikely to be able to influence the special education placement of students with many impairments, such as a visual impairment or traumatic brain injury. I consider the likelihood of being in one of two broad categories of diagnoses – those that are likely to be relatively malleable and those that seem especially

¹² While taking an alternate test might improve the score of a student who would struggle to demonstrate their knowledge on a standard test, the most common of these tests, the NCEXTEND2, evaluated students relative to grade-level standards. For example, the reading form “uses shorter reading selections, simplified language, and fewer test items and item responses (foils/answer choices) to assess students on grade-level content” (North Carolina Public Schools, 2009, p 5). The cut scores between achievement levels were selected through a similar procedure for both the NCEXCEND2 and the regular end of grade tests. First, a group of students who met the eligibility criteria piloted the tests. Then the teachers of these students were asked to use their knowledge of the students’ classroom performance to categorize them into achievement levels. Test makers noted the percentage of students expected to score in each achievement level, and set cut scores accordingly. That is, if the teachers reported that 15 percent of the 4th grade students tested were in achievement level 1 the test makers set the cut off between levels 1 and 2 such that the lowest 15 percent of scores were in level 1 (North Carolina Public Schools, 2009). Cut scores were then reviewed and approved by a panel of policy makers and stakeholders. To the extent that this assumption is incorrect my results would be biased towards finding positive effects of special education, suggesting that its true negative effects are even larger than estimated.

difficult for schools to alter. Malleable impairments are speech and language impairments, learning disabilities, emotional and behavioral disorders, and other impairments (which includes ADHD). Non-malleable impairments are autism, intellectual disability, developmental disabilities, sensory disabilities, traumatic brain injury, orthopedic impairments, and multiple disabilities. I focus on malleable diagnoses defined this way in my main analysis. Estimates that include students with autism in the malleable group appear in Appendix Table A.6 and are similar to my main results. I also use the non-malleable diagnoses to conduct a falsification test.

Table 1. Distribution of Diagnoses in Special Education

	Percent of Special Education	
	Percent of Students	Students
Autism	0.7	4.9
Deaf-Blindness	0.0	0.0
Developmental Delay	0.0	0.0
Emotional Disturbance	0.6	4.7
Hearing Impairment	0.1	1.1
Intellectual Disability	1.9	14.1
Multiple Disabilities	0.1	1.1
Orthopedic Impairment	0.1	0.5
Other Health Impairment	2.6	19.4
Specific Learning Disability	5.9	44.8
Speech or Language Impairment	1.2	8.8
Traumatic Brain Injury	0.0	0.2
Visual Impairment	0.0	0.3
Total	13.2	100

Notes: Table reports the percent in special education by diagnosis for the sample of students in North Carolina Title I schools in grades 4-8 in years 2006-7 – 2010-11. Diagnoses shaded in grey (emotional disturbance, other health impairment, specific learning disability, speech or language impairment) are included in the “malleable impairment” category.

Data from the alternative tests are available beginning in 2006, so I begin my analysis in 2007 to have at least one previous year of data for students taking the alternative tests. This restriction also allows me to analyze a consistent policy environment, as North Carolina

began using student gains in its AYP calculation in 2005-2006. I also exclude third-grade students, as most do not have a prior year test score.

I remove from the analysis sample students who are missing information on current special education status, current or previous standardized test scores, or the performance of their school and subgroups in the past.¹³ I loosen the requirement to have current-year scores when considering whether schools altered the testing population in response to incentives. Students with incomplete data are only included in the sample for the years and subjects for which complete information is available, resulting in an unbalanced panel. I investigate the relationship between incentives and attrition in section 5.4. In baseline specifications, I drop all those who ever appear in the data with a non-malleable diagnosis, so as not to confuse changes of diagnosis with movement in and out of special education. This restriction is altered when considering the effects of incentives on having a non-malleable diagnosis, and as a robustness check in Appendix Table A.2. Results are not sensitive to the exclusion of students who had a non-malleable diagnosis at some point in time.

My main sample, described in Table 2, comprises about 1.3 million student-year observations, representing about 700,000 students in seventeen hundred schools.¹⁴ About 10.5 percent of the students in my sample were in special education. While about 13 percent of students in North Carolina were in special education, my main sample excludes those who had a non-malleable diagnosis at any point in time, decreasing the percent in special

¹³ Of students in the grades and years considered, 8.9 percent are missing prior-year test scores. This includes students who are in their first year in North Carolina Public Schools. Among those with prior-year test scores, 1.4 percent are missing information on prior school performance. Of those with data on prior test scores and prior school performance 0.4 percent are missing information on school Title I status.

¹⁴ Descriptive statistics for those who ever had a non-malleable diagnosis appear in Table A.3.

education. About half were female, and 57 percent were eligible for free or reduced-price lunch. Slightly less than half the sample identified as White/Caucasian, 31 percent as Black, and 12 percent as Hispanic. Most had passed their standardized test in the previous year, 69 percent in reading and 73 percent in math. Twenty two percent were in schools that had failed to achieve AYP thresholds in math for at least one subgroup in the previous year, and 18 percent were in schools that had failed to do so in reading. Students in special education were significantly less likely to have passed their test in the previous year and more likely to be low-income. This demonstrates the disadvantaged nature of the special education population, which is part of the empirical challenge of identifying the causal effect of placement.

Table 2. Descriptive Statistics for the Main Analysis Sample

	All		Not Special Education		Special Education	
	Mean	SD	Mean	SD	Mean	SD
Special Education	0.105	0.306				
Prior Pass Reading	0.692	0.462	0.730	0.444	0.364	0.481
Prior Pass Math	0.732	0.443	0.764	0.425	0.463	0.499
Native American	0.023	0.151	0.023	0.150	0.024	0.154
Asian	0.017	0.130	0.018	0.135	0.007	0.085
Hispanic	0.124	0.329	0.126	0.332	0.104	0.306
Black	0.311	0.463	0.307	0.461	0.351	0.477
White	0.488	0.500	0.489	0.500	0.476	0.499
Other	0.037	0.188	0.037	0.188	0.037	0.190
Female	0.497	0.500	0.516	0.500	0.331	0.470
Free or Reduced-Price Lunch	0.570	0.495	0.556	0.497	0.691	0.462
School failed in math	0.221	0.415	0.220	0.414	0.222	0.416
School failed in reading	0.184	0.387	0.184	0.388	0.184	0.387
Math Score	-0.100	0.946	-0.024	0.923	-0.743	0.889
Prior math score	-0.109	0.948	-0.030	0.923	-0.779	0.894
Reading Score	-0.105	0.962	-0.022	0.924	-0.814	0.979
Prior reading score	-0.107	0.965	-0.015	0.922	-0.896	0.959
N	1,298,002		1,161,922		136,080	

Notes: This table presents descriptive information on the main analysis sample. Test scores are in standard deviation units.

5. School Responses to Accountability Incentives

5.1 Method

I first examine school responses to accountability incentives, then consider the effect of special education placement on student outcomes. I assume that schools make decisions about special education placements at least once a school year. This is consistent with US Department of Education regulations that require IEPs to be reviewed every 12 months or more often if necessary (US Department of Education, 2000). When making these choices, they may consider a wide array of information about their students but only take AYP incentives for the **current** year into account. This assumption would be violated if a school tried to slow its improvement this year in order to make improving next year easier. While it is likely that schools would want to plan ahead, it seems less likely that they would be able to do so effectively. This assumption allows me to consider a static model in which schools respond to current incentives. If it is incorrect my estimates will not reflect all responses to accountability incentives but would still reflect current year responses to current year incentives - a relevant parameter.

I address the question of how NCLB incentives altered disability classifications by testing two main hypotheses about school responses to incentives as outlined in Section 2.4. First, schools that are at risk of failing to make AYP may use special education as a way to target extra services to “bubble students” whose passing status could reasonably be changed.

$$(1) SE_{igt} = \beta_1 X_{igt} + \sum_{s=r,m} [\beta_{2s} Incentive_{igjst} + \sum_{a=2,3} [\beta_{3sa} Bubble_{igjsat} + \beta_{4sa} Incentive_{igjst} * Bubble_{igjsat}]] + \gamma_{gt} + \sigma_j + \varepsilon_{igt}$$

In Equation (1), an indicator for whether student i in grade g in school j in year t is in special education (SE_{igt}) is a function of observed characteristics of student i (X_{igt}) as well as whether student i is a bubble student ($Bubble_{igjst}$) just above the passing threshold ($a = 3$) or just below ($a = 2$) in reading ($s = r$) or math ($s = m$), whether the school has an accountability incentive to ensure that student i passes ($Incentive_{igjst}$) as defined in the next paragraph, and an interaction between those final terms for each subject. All models include year by grade fixed effects (γ_{gt}). Main estimates include school fixed effects (σ_j); estimates without school fixed effects are available from the author upon request. Student characteristics include prior scores in math and reading as well as indicators for lagged special education status, LEP status, free or reduced-price lunch eligibility, and the racial and ethnic categories used by NCLB.

I consider a school to have an incentive to improve the likelihood of student i passing if the school failed to make AYP in the previous year for any subgroup – not including the SWD group– into which student i falls and would face sanctions for future AYP failures. For example, if student i is economically disadvantaged and Hispanic the school would have an incentive to ensure that the student passes if the school was at risk of failing AYP for all students, Hispanic students, or economically disadvantaged students, and would face consequences for doing so. As discussed in more detail in section 2.2, some schools that did not achieve AYP requirements avoided AYP failure status. Schools with these types of failures in the past year would not receive immediate sanctions if they failed in the current year.

I define bubble students in two ways. First, I consider that schools may target their actions bluntly based on the achievement levels in which students scored in the previous year. To do this I define bubble status as having scored in achievement level 2 or

achievement level 3 in the previous year. Because schools may treat students who previously passed differently from those who previously failed, I include separate terms for being in level 2 or level 3.

Second, I consider that schools may target their actions more precisely to students particularly close to the passing threshold. In doing so, I incorporate North Carolina's gain score model, under which any student who was on a trajectory to reach proficiency within 4 years of their first test could be counted as passing for AYP targets. I define an inverse measure of the amount by which a student's score would need to rise or fall in order for them to just count as passing for AYP determinations. I begin by defining the student's distance from counting as passing. For students who passed and students who failed but are in at least 7th grade, this is the absolute difference between the student's score and the test's cut score. Students in sixth grade or lower who failed the previous year can count as passing in the current year if they improve enough to be on a trajectory to be proficient by seventh grade. I approximate this needed improvement as their distance from the cut score divided by the number of years left before seventh grade. I then use the distance from counting as passing to construct an inverse distance measure standardized across grades, years, and subjects. To do this I calculate the largest distance to counting as passing for each grade-year-subject combination. The standardized measure of the inverse distance to counting as passing is the largest distance for the grade-year-subject minus the student's distance, divided by the largest distance for the grade-year-subject. In my models, I test whether this measure of distance matters for either students who previously scored in level 2 or in level 3.

If schools use special education to target services to almost-passing (just-passing) students, I would expect to find positive coefficients on β_{4r2} and β_{4m2} (β_{4r3} and β_{4m3}) when using the simple definitions of bubble group membership. To the extent that schools focus

on those very close to passing even among level 2 (level 3) students, I would also find positive values when defining bubble status based on the inverse distance from counting as passing for level 2 (level 3) students.

Second, schools that are at risk of failing to make AYP for the SWD subgroup could attempt to select their special education population to be relatively high performing. I test for this possibility using the following model:

$$(2) SE_{igt} = \beta_1 X_{igt} + \sum_{s=r,m} [\beta_{2s} \widehat{Pass}_{igt} + \beta_{3s} MarginalSWD_{jst} + \beta_{4s} \widehat{Pass}_{igt} * MarginalSWD_{jst}] + \gamma_{gt} + \sigma_j + \varepsilon_{igt} .$$

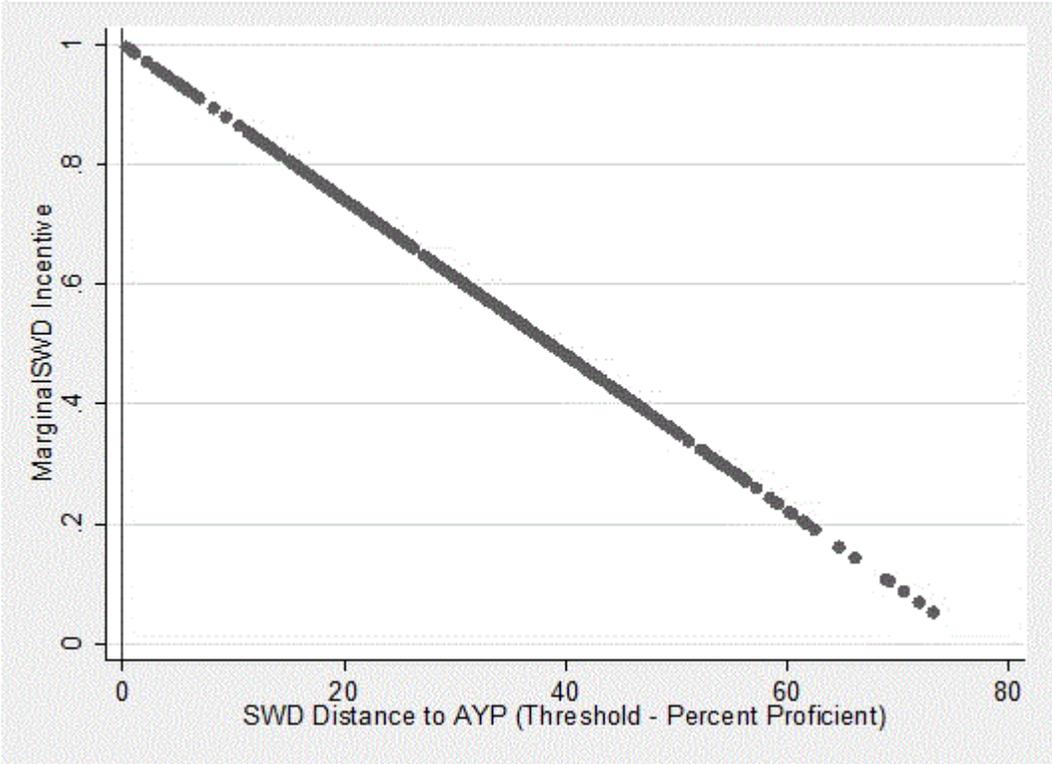
The special education status of student i still depends on student characteristics and year-by-grade fixed effects. Now, the school's incentive is an interaction between the student's predicted performance (\widehat{Pass}_{igt}) in each subject and whether the school is at the margin of failing to make AYP in that subject for the SWD subgroup ($MarginalSWD_{jst}$). I assume that a school's best prediction about a student's performance this year is their score last year (or, alternately, whether they were proficient last year). To define to what extent schools are at the margin of failing to make AYP in a subject for the SWD subgroup, I create an inverse measure of the amount a school would have to improve their performance to meet AYP.

$$(3) MarginalSWD_{jst} = \left(1 - \left(\frac{Threshold_{jst-1} - PercentProfSWD_{jst-1}}{Threshold_{jst-1}} \right) \right) * FailedSWD_{jst-1}$$

In Equation (3), the degree to which school j 's SWD subgroup is marginal to passing in subject s and year t ($MarginalSWD_{jst}$) is defined based on the schools' percent proficient in the previous year ($PercentProfSWD_{jst-1}$), the AYP threshold for that subject and year ($Threshold_{jst-1}$), and an indicator for whether the school had failed to make AYP for the SWD group in the previous year ($FailedSWD_{jst-1}$). Figure 2 illustrates the MarginalSWD

measure for a sample of schools that had failed to make AYP in the previous year. The measure takes on a value of 0 for schools in which no SWD students passed in the previous year and climbs linearly with the percent of SWD students passing until reaching a value of nearly 1 for schools just below the AYP threshold. The measure is 0 for schools in which the SWD group achieved AYP, either by having a passing rate at or over the threshold, or through one of the alternate calculations discussed earlier.

Figure 2. SWD Incentive Instrument



Notes: This figure depicts the SWD incentive instrument as described in Equation (3).

In reality, many schools that fail in one group fail in more than one – roughly half of schools that failed in the SWD subgroup also failed in another group and most schools that fail in another group fail in the SWD subgroup. As a result, many schools are faced with both incentives simultaneously. Some of the students in these schools will also be the targets of both incentives; consider a student who just passed and whose school failed to make AYP

both for the SWD group and a demographic subgroup of which the student is a member. For this reason, I model them together as in equation (4) below.

$$(4) SE_{igt} = \beta_1 X_{igt} + \sum_{s=r,m} [\beta_{2s} Incentive_{igjst} + \sum_{a=2,3} [\beta_{3sa} Bubble_{igjsat} + \beta_{4sa} Incentive_{igjst} * Bubble_{igjsat}]] + \beta_{5s} \widehat{Score}_{igjst} + \beta_{6s} MarginalSWD_{jst} + \beta_{7s} \widehat{Score}_{igjst} MarginalSWD_{jst}] + \gamma_{gt} + \sigma_j + \varepsilon_{igt}$$

This model identifies causal effects of accountability-related incentives on placement under two assumptions. For the first hypothesis, the assumption is that placement differences between bubble and non-bubble students when schools do not have an AYP incentive to improve the students' likelihood of passing reflect the differences by bubble status that would exist for students whose schools have an AYP incentive to improve their likelihood of passing, in the absence of that incentive. Suppose for a given year and subject NCLB required 84 percent proficiency for all groups. The parallel trends assumption would be violated if, in the absence of AYP incentives, the change in placement that a bubble student experiences when one of the subgroups to which they belong goes from having at least 84 percent proficiency to less than 84 percent proficiency was different from the change experienced by a non-bubble student.

For the second hypothesis, the assumption is that differences in placement by prior test score would not vary with the SWD group's passing rate relative to the AYP threshold in the absence of AYP incentives. This would be violated if, in the absence of AYP incentives, the change in placement a previously passing student would experience if their school's SWD group went from above 84 percent proficiency to below was different from that of a previously failing student.

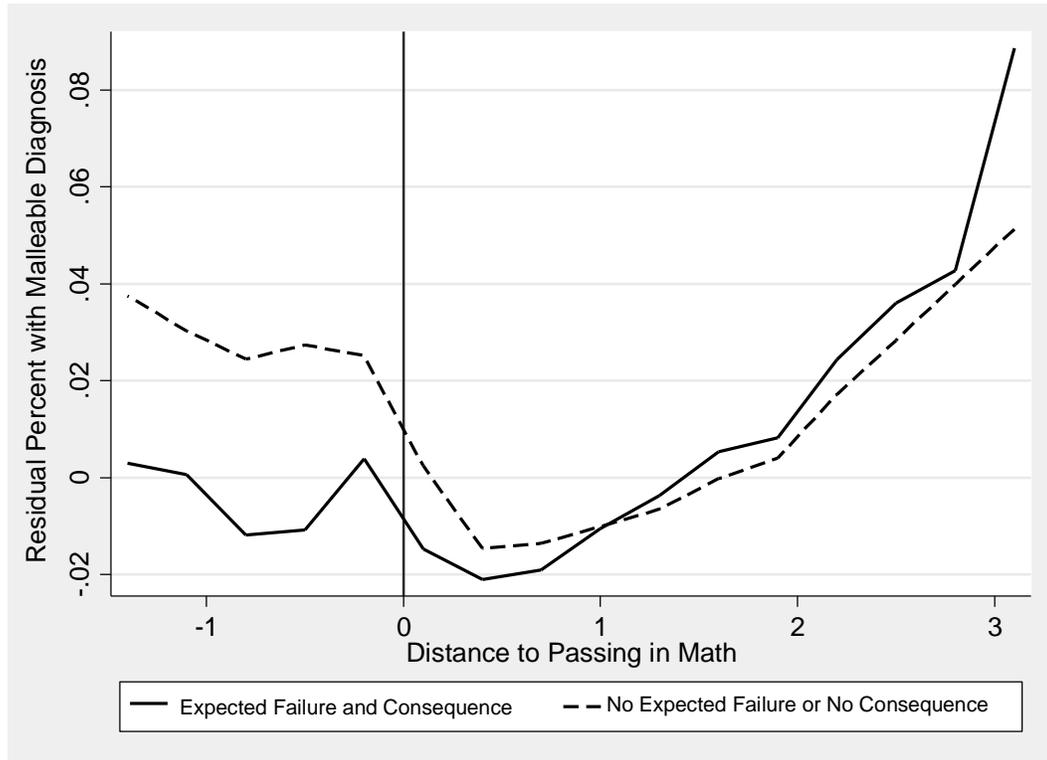
Students and parents have their own incentives surrounding special education, and schools face other sources of pressure. However, most student and parent incentives do not change around student passing thresholds, and those that do should not change around schools' AYP thresholds. That is, a student's parents may want them to be in special education to improve their performance, and this will appear in coefficients on prior score or demographics. Some of those families will probably push harder for placement if their student is struggling, or perhaps if they appear to be almost doing "well enough" but need a slight boost. Schools may similarly use scores to identify struggling students. Both of these responses will appear in the coefficient on expected score or being a "bubble" student. In the absence of accountability incentives, these reasons for special education classification do not change when the school is at risk of failing to make AYP or when the student is important to the school's effort to do so. Similarly, in a world without NCLB, schools' identification of struggling students by their score should not depend on the student's subgroup membership or the SWD group's performance. Thus, the interaction terms that identify school responses to incentives and form my instruments should reflect only school responses as a result of NCLB incentives.

5.2 Placement Responses to AYP Incentives

Figure 3 illustrates the residual percentage of students in the main analysis sample who had a malleable diagnosis after controlling for demographics, prior score, and year-by-grade fixed effects, by the student's distance to a passing score in math the previous year. The first series, marked with a solid line, includes all those whose school expected to fail AYP for a group of which that student was a member, and the school would face consequences for such a failure ($Incentive_{ijmt} = 1$). The second series, marked with a dashed line, includes those whose school either did not expect to fail AYP for any group of which that

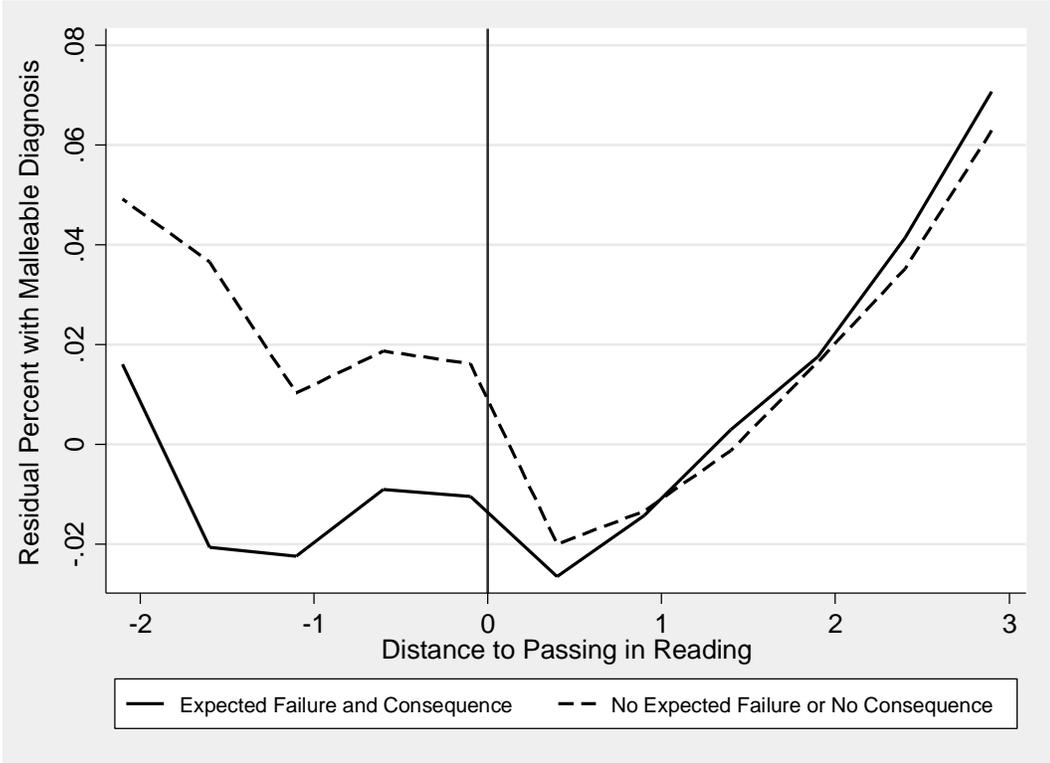
student was a member or did not face sanctions for such a failure ($Incentive_{ijmt} = 0$). Most students (about 69 percent) whose school had previously failed to make AYP for a group of which they were a member also failed to make AYP for the SWD group, so had positive values of $MarginalSWD_{jst}$. Nearly all students whose schools did not have an incentive to improve their likelihood of passing also did not have an incentive to improve the performance of the SWD group (97 percent). As a result the figure illustrates responses to both incentives simultaneously.

Figure 3. Residual Percent with Malleable Diagnosis by Distance to Passing Score in Math and Accountability Incentives



Notes: This figure displays the residual percent of students with a malleable diagnosis, after controlling for prior score and demographics, by the student's distance from the passing threshold measured in standard deviations. The first series, marked with the solid line, includes students whose school expected to fail AYP for at least one group of which the student was a member, and would potentially face consequences for doing so. The second series, marked with the dashed line, includes all other students.

Figure 4. Residual Percent with Malleable Diagnosis by Distance to Passing Score in Reading and Accountability Incentives



Notes: This figure displays the residual percent of students with a malleable diagnosis, after controlling for prior score and demographics, by the student’s distance from the passing threshold measured in standard deviations. The first series, marked with the solid line, includes students whose school expected to fail AYP for at least one group of which the student was a member, and would potentially face consequences for doing so. The second series, marked with the dashed line, includes all other students.

Students who previously received a failing score or just passed were more likely to be in special education with a malleable diagnosis if their school did not have an incentive to improve their performance. Those who had previously passed by more than about a standard deviation were about as likely to be in special education regardless of whether their school had an incentive to improve their performance. There is a similar pattern in reading, illustrated in Figure 4. If hypothesis 2 is correct we would expect the relatively likelihood of being in special education for previously-passing versus previously failing students to be higher in schools with incentives to improve the performance of the SWD subgroup. Figures

3 and 4 suggest that this is the case, and that the lion's share of the selection takes place through discouraging special education for previously-failing students, rather than encouraging special education for previously-passing students.

If schools use special education to target supports and services to bubble students when the school would benefit from their achieving a passing score, as suggested by hypothesis 1, we would expect the presence of an AYP incentive to increase residual malleable diagnoses close to the passing threshold. There is a noticeable bump in malleable diagnoses in Figure 3, peaking around a quarter of a standard deviation below the passing threshold. The pattern appears more dramatic for those whose schools had an AYP incentive to improve their likelihood of passing, but is also present for those without such an incentive. A broadly similar pattern holds in reading, as illustrated in Figure 4.

I now turn to a more systematic examination of how schools respond to accountability incentives using Equation (4). The next several tables are structured similarly. Each column displays estimates from a single regression. In the odd columns, "bubble" status is defined using binary indicators of achievement level, while the even columns test whether the distance from the cut score matters for students in levels 2 or 3, as detailed in the previous section. In columns 1 and 2, I test whether schools select students based on prior passing status when trying to improve the performance of the SWD subgroup; columns 3 and 4 use prior score. Student demographics, year-by-grade fixed effects, and school fixed effects are included in all regressions. Estimates without school fixed effects are available from the author upon request.

Table 3. Effect of Accountability Incentives on Special Education Placement

	(1)	(2)	(3)	(4)
Reading Level 2*Incentive	0.0142** (0.00302)		0.00930** (0.00283)	
Math Level 2*Incentive	-0.0115** (0.00353)		-0.0133** (0.00300)	
Reading Level 3*Incentive	0.0252** (0.00285)		0.0239** (0.00272)	
Math Level 3*Incentive	-0.000835 (0.00193)		-0.000850 (0.00222)	
Reading Level 2*Distance*Incentive		0.0143** (0.00319)		0.0101** (0.00299)
Math Level 2*Distance*Incentive		-0.0139** (0.00388)		-0.0140** (0.00335)
Reading Level 3*Distance*Incentive		0.0277** (0.00323)		0.0284** (0.00312)
Math Level 3*Distance*Incentive		-0.00352 (0.00233)		-0.000355 (0.00277)
Reading Prior Pass*SWD Incentive	0.0221** (0.00463)	0.0223** (0.00469)		
Math Prior Pass*SWD Incentive	0.0125 (0.00726)	0.0139 (0.00756)		
Reading Prior Score*SWD Incentive			0.0110** (0.00254)	0.0115** (0.00258)
Math Prior Score*SWD Incentive			0.00682** (0.00262)	0.00706** (0.00273)
N	1199737	1199737	1199737	1199737

Notes: This table displays results from 4 linear probability models following Equation (4), one in each column, in which being in special education with a malleable diagnosis is the dependent variable. Each model includes demographic controls, year-by-grade fixed effects, school-level fixed effects, and the main effects of school incentives and prior student performance. Standard errors in parentheses are clustered at the school level. * denotes significance at the 0.05 level, ** at the 0.01 level.

Table 3 presents the effects of NCLB incentives on the likelihood of being in special education with a malleable diagnosis. I will start by discussing the second hypothesis, that schools select the special education population to be relatively high-performing when the school would benefit from improving the performance of the SWD group. Students who had previously passed in reading were 2.2 percentage points more likely to be in special

education when their school had an incentive to improve the performance of the SWD group in reading, as shown in columns 1 and 2. I find no significant evidence of selection based on math performance with this specification. Considering score rather than passing status, as in columns 3 and 4, suggests that a one standard deviation higher prior reading score increased the likelihood of being in special education by about 1 percentage point when the student's school had just failed to achieve AYP for the SWD group. A math score that was one standard deviation higher increased the likelihood that a student would be in special education by 0.7 percentage points when the student's school had just failed to achieve AYP in math for the SWD group.

Next, I consider evidence of the first hypothesis. I find that schools encouraged special education placement for students who were close to the passing threshold in reading, whether above or below. In column 1, when bubble status is measured by having previously scored in level 2 or 3, a student who had previously scored in level 2 (level 3) in reading would be 1.4 (2.5) percentage points more likely to be in special education if their school would benefit from their passing. These estimates are robust to the way selection of high-performing students is parameterized. I also find evidence that schools targeted those closer to the passing threshold more strongly than those farther away, as shown in columns 2 and 4.

Schools appear to have discouraged students who had scored in level 2 in math from special education when the school would benefit from that student passing, and not changed their treatment of level 3 students in math. This may reflect schools believing that there are other options available to raise the math scores of almost-passing students. It could also be a manifestation of school attempts to control the size of the special education population. To investigate this, it is useful to compare schools that had at least 12.5 percent of their student

body in special education, and would not receive additional state aid to support further special education placements, with those schools with a smaller proportion in special education. As shown in Table A.1, schools with at least 12.5 percent of their students in special education appear to discourage placement more strongly for students who had scored in level 2 in math than do schools with a smaller proportion of students in special education. Schools with more than 12.5 percent of their students in special education also encourage special education less strongly for students who scored in level 2 in reading, in comparison to schools with smaller special education populations.

5.3 Heterogeneity

One way of verifying that the estimates from Table 3 reflect school responses to AYP incentives is to compare these reactions across groups that faced stronger and weaker incentives. If estimates reflect a causal relationship rather than omitted variables those who faced stronger incentives should have exhibited larger reactions, or at least not smaller.

While a school's incentives surrounding a student do not necessarily depend on that student's underlying impairment, a school's ability to influence whether the student is in special education does. Schools should have much less influence on diagnoses for which it would be difficult to not place a student in special education – say a student who is blind or uses a wheelchair – than on the relatively malleable diagnoses I consider in my main analyses. Table 4 displays the results of relaxing the sample restriction that excluded those who had ever had a non-malleable diagnosis and estimating the effect of incentives on non-malleable diagnosis. I find some evidence that students who had just passed in the previous year were less likely to have a non-malleable diagnosis when their school would benefit from their passing. However, these effects are quite small in comparison to the results in Table 3, and no other coefficients are significant. This does however highlight the fact that some of

the diagnoses I am classifying as non-malleable can be influenced by schools, just to a lesser extent than the malleable diagnoses.

Table 4. Effect of Accountability Incentives on Having a Non-Malleable Diagnosis

	(1)	(2)	(3)	(4)
Reading Level 2*Incentive	0.00113 (0.00184)		0.000556 (0.00178)	
Math Level 2*Incentive	0.00035 (0.00222)		-0.00121 (0.00202)	
Reading Level 3*Incentive	-0.00205 (0.00176)		-0.00390* (0.00170)	
Math Level 3*Incentive	-0.00053 (0.00153)		-0.00365* (0.00169)	
Reading Level 2*Distance*Incentive		0.000717 (0.00194)		-0.0000361 (0.00189)
Math Level 2*Distance*Incentive		-0.00099 (0.00238)		-0.000853 (0.00221)
Reading Level 3*Distance*Incentive		-0.00282 (0.00200)		-0.00476* (0.00193)
Math Level 3*Distance*Incentive		-0.00064 (0.00175)		-0.00282 (0.00201)
Reading Prior Pass*SWD Incentive	-0.00039 (0.00257)	-0.00041 (0.00259)		
Math Prior Pass*SWD Incentive	-0.00357 (0.00392)	-0.00475 (0.00399)		
Reading Prior Score*SWD Incentive			0.00253 (0.00147)	0.00261 (0.00197)
Math Prior Score*SWD Incentive			-0.00184 (0.00156)	-0.00189 (0.00162)
N	1237846	1237846	1237846	1237846

Notes: This table displays results from 4 linear probability models following Equation (4), one in each column, in which being in special education with a non-malleable diagnosis is the dependent variable. Each model includes demographic controls, year-by-grade fixed effects, school-level fixed effects, and the main effects of school incentives and prior student performance. Standard errors in parentheses are clustered at the school level. * denotes significance at the 0.05 level, ** at the 0.01 level.

I also examine reactions by schools that did not receive Title I funding. These schools were still required to test their students in accordance with NCLB mandates, and performance was reported publicly. As such they did have incentives to perform well according to the NCLB metrics, but these incentives were much lower than for Title I schools,

which faced the possibility of sanctions. Unfortunately for this analysis, very few non-Title I schools failed to achieve AYP, so there is little variation in the data, resulting in imprecise estimates. These results appear in Table A4, and suggest that schools that did not receive Title I funding did not use special education to target services to level 3 students whose passing would benefit the school. There is some evidence of targeting away from level 2 students, particularly in reading, and selection against prior-passing students in reading, both of which are somewhat puzzling. They may reflect school attempts to keep the special education population below the 12.5 percent funding threshold, while using placement to respond to other priorities. Non-Title-I schools do appear to have selected students based on their math performance when the school expected its SWD group to fail. These point estimates are larger than those for Title I schools, but are quite imprecise.

5.4 Test Taking and Selection

One of the primary innovations of NCLB was its testing requirement – schools were required to have at least 95 percent of students contributing scores, both overall and in each accountable subgroup. This drastically decreased the scope for schools to select the tested population but did not eliminate it entirely. Schools could still potentially do more to encourage test-day attendance for some students than others, and those with especially good attendance might be able to actively discourage some particularly low-scoring students. Selection of the test taking population would be another way for schools to respond to accountability incentives that would not appear in my main analysis. It also would limit my ability to use these incentives as instruments for special education placement in the next section. Suppose special education had no effect on student achievement, but the same accountability pressures that influenced selection into special education drove schools to

change the tested population. In this case it would be possible to find effects of special education – in either direction depending on how selection into testing took place.

I investigate whether accountability incentives predict the likelihood a student appears in the data with a valid test score. Results appear in Table 5. I find no evidence that schools respond to AYP incentives by altering the tested population. This suggests that my analysis of the effect of special education on achievement outcomes is not subject to bias due to sample selection, something that previous research on the subject likely suffered from and was not able to analyze directly. It also suggests that the combination of NCLB's testing requirements and the introduction of alternate tests succeeded in discouraging schools from excluding their special education students from testing.

Table 5 Effect of Accountability Incentives on Not Having a Valid Score

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Reading	Reading	Reading	Reading	Math	Math	Math	Math
Reading Level 2*Incentive	0.00157 (0.00418)		0.00217 (0.00416)		0.00150 (0.00419)		0.00212 (0.00417)	
Math Level 2*Incentive	0.00535 (0.00398)		0.00368 (0.00353)		0.00542 (0.00398)		0.00373 (0.00352)	
Reading Level 3*Incentive	-0.00221 (0.00391)		-0.00368 (0.00382)		-0.00230 (0.00391)		-0.00379 (0.00383)	
Math Level 3*Incentive	-0.00194 (0.00276)		-0.000234 (0.00276)		-0.00190 (0.00276)		-0.000180 (0.00275)	
Reading Level 2*Distance*Incentive		-0.00234 (0.00447)		-0.00398 (0.00436)		-0.00239 (0.00448)		-0.00406 (0.00436)
Math Level 2*Distance*Incentive		-0.00173 (0.00321)		0.000448 (0.00324)		-0.00170 (0.00320)		0.000519 (0.00323)
Reading Level 3*Distance*Incentive		0.00215 (0.00466)		0.00281 (0.00464)		0.00212 (0.00466)		0.00280 (0.00465)
Math Level 3*Distance*Incentive		0.00662 (0.00438)		0.00483 (0.00393)		0.00673 (0.00437)		0.00491 (0.00392)
Reading Prior Pass*SWD Incentive	-0.00548 (0.00448)	-0.00550 (0.00451)			-0.00561 (0.00448)	-0.00564 (0.00451)		
Math Prior Pass*SWD Incentive	0.00947 (0.00653)	0.00950 (0.00646)			0.00960 (0.00653)	0.00967 (0.00647)		
Reading Prior Score*SWD Incentive			0.000142 (0.00226)	0.0000902 (0.00226)			0.0000824 (0.00226)	0.0000268 (0.00226)
Math Prior Score*SWD Incentive			0.00156 (0.00227)	0.00161 (0.00228)			0.00160 (0.00227)	0.00166 (0.00229)
N	1199737	1199737	1199737	1199737	1199737	1199737	1199737	1199737

Notes: This table displays results from 8 linear probability models following Equation (4), one in each column, in which not having a valid reading or math score is the dependent variable. Each model includes demographic controls, year-by-grade fixed effects, school-level fixed effects, and the main effects of school incentives and prior student performance. Standard errors in parentheses are clustered at the school level. * denotes significance at the 0.05 level, ** at the 0.01 level.

6 Effects of Special Education on Student Achievement

6.1 Method

Special education is not randomly assigned, and students who receive special education are systematically different from those who do not. As a result, simple comparisons of the outcomes of students in and out of special education would not reflect the causal effect of placement. To overcome this problem I use incentives from the previous section as instruments for special education. In order for the incentives to be valid instruments they must obey the exclusion restriction. That is, they must not affect outcomes through some mechanism other than special education placement. I do not expect this the exclusion restriction to hold in the case of the incentives from the first hypothesis, as prior work suggests that schools are able to use other efforts to target resources to students who they wish to pass (Reback 2008). For this reason, I do not use the incentives from the first hypothesis as instruments.

The incentives from the second hypothesis make more suitable instruments. For these instruments, the exclusion restriction holds as long as school incentives to select the SWD population do not affect students' test scores differentially by score, except through changes to special education placement. One threat to this assumption is the possibility that schools expend extra effort on selected students to ensure that they pass. The restriction would not be violated if the school is making efforts to improve the test performance of students in special education in general as a result of accountability pressure for that group. To the extent that special education under accountability pressure is different from that without, it could limit generalizability. However, my results would still apply to any situation where schools make a particular attempt to increase the percentage of students in special education achieving proficient scores. Similarly, if some resources or opportunities are not

provided to students in special education, perhaps due to time or scheduling constraints, this would not bias my results. Rather, it would mean that a move into special education entailed not only the addition of services detailed in the student's IEP but a loss of other services provided to students who were not in special education. If for some reason this withdrawal of resources only took place when the school was under accountability pressure the result might not fully generalize to environments where special education was an additive service rather than the exchange of one set of services for another. It is not clear why this would be true, and even then the exclusion restriction would not be violated. Finally, it is possible that students who are selected into special education based on scoring well in one year experience a decrease in score in the next. This reversion to the mean would only confound my estimates if it somehow occurred for students in schools with an incentive to improve the performance of the SWD group and not for those in schools without that incentive or vice versa. There is no reason to believe that this is the case.

It is also necessary for the instruments to satisfy a monotonicity assumption. This would be violated if some students I have labelled as incentive targets were not seen as such by their schools and instead were discouraged from being in special education in order to "make space" for others. This would require a school to believe that students who failed in the previous year were more likely to pass in the current year than those who passed in the previous year, which seems extremely unlikely.

That is (4) serves as the first stage of a model with the second stage:

$$\begin{aligned}
(5) \text{ Score}_{igt} &= \beta_1 X_{igt} \\
&+ \sum_{s=r,m} [\beta_{2s} \text{ Incentive}_{igjst} + \sum_{a=2,3} [\beta_{3sa} \text{ Bubble}_{igjsat} \\
&+ \beta_{4sa} \text{ Incentive}_{igjst} * \text{ Bubble}_{igjsat}]] + \beta_{5s} \widehat{\text{Score}}_{igjst} \\
&+ \beta_{6s} \text{ MarginalSWD}_{jst}] + \beta_7 \widehat{SE}_{igt} + \gamma_{gt} + \sigma_j + \varepsilon_{igt}
\end{aligned}$$

In Equation (5) Score_{igt} is the relevant current-year test score or other achievement outcome. Special education status (\widehat{SE}_{igt}) is a simple indicator for whether the student is in special education with a malleable diagnosis in the current year. The model also includes student characteristics as well as year-by-grade fixed effects and school fixed effects. Estimates without school fixed effects are available from the author upon request.

6.2 Effects of Special Education on Achievement

I use Equation (5) to estimate the effect of being placed in special education on a marginal student's same-year achievement. Incentives to select the SWD group to be relatively high performing in reading form strong instruments, as shown in Table 6. I do not find evidence that special education has an effect on achievement for students whose placement is altered by incentives to select the SWD population to have a higher passing rate in reading. These estimates are noisy, so I cannot rule out large effects in either direction. Math incentives form weaker instruments; only those based on prior score are strong. However, these estimates suggest that special education hurts math achievement for those students whose special education placement is altered to improve the math achievement of their school's SWD group. Being placed in special education lowers math scores by about 1.2 standard deviations for this group of students. The point estimates for reading scores for this group are also negative, though smaller and not significant at conventional levels. A 1.2 standard deviation effect on test score is extremely large, roughly equivalent to falling from

the 75th percentile to the 25th. About 4 percent of students in my sample experience a year-to-year change at least this large. However a 1.2 standard deviation effect could reflect a 0.6 standard deviation fall from a student who would otherwise have experienced a 0.6 standard deviation gain. About 28 percent of my sample experiences a year-to-year score change of at least 0.6 standard deviations.

Table 6. Effect of Special Education on Student Achievement

Panel 1: Reading score, reading instrument				
	(1)	(2)	(3)	(4)
Special Education	0.275 (0.389)	0.577 (0.344)	0.291 (0.386)	0.581 (0.329)
F-Statistic	25.35	24.78	25.39	26.03
N	1199505	1199505	1199505	1199505
Panel 2: Math score, reading instrument				
Special Education	0.0451 (0.343)	0.171 (0.320)	0.0107 (0.334)	0.249 (0.309)
F-Statistic	25.33	24.77	25.37	26.02
N	1199496	1199496	1199496	1199496
Panel 3: Reading score, math instrument				
Special Education	-1.827 (1.034)	-0.770 (0.484)	-1.714 (0.947)	-0.685 (0.470)
F-Statistic	5.778	14.28	6.388	14.09
Panel 4: Math score, math instrument				
Special Education	-1.501 (0.941)	-1.289* (0.580)	-1.623 (0.917)	-1.129* (0.555)
F-Statistic	5.777	14.28	6.388	14.09
Instruments:				
Prior Pass*SWD Incentive	Y		Y	
Prior Score*SWD Incentive		Y		Y
Controls:				
Levels*Incentive	Y	Y		
Levels*Distance*Incentive			Y	Y

Notes: This table displays results from 16 linear IV models following Equation (5), one in each column and panel, in which being in special education with a malleable diagnosis is instrumented by selection incentives as noted and math or reading z-score is the dependent variable. All specifications include demographic controls, year-by-grade fixed effects, and school-level fixed effects, and I control for hypothesis 1 incentives as noted. Standard errors in parentheses are clustered at the school level. Kleibergen-Papp F-statistics from the first stage are reported. * denotes significance at the 0.05 level, ** at the 0.01 level.

The fact that special education can hurt the achievement of some students is surprising given prior findings of the effect of special education on achievement. However, my results are consistent with the findings by Setren (2016) that students who had a special education placement before entering a charter school saw gains similar to those of their non-special education classmates despite losing special education designations at a fairly high rate. The differences between my estimates and those in Hanushek, Kain, and Rivkin (2002) could be driven by differences in settings – location, time period, grades, etc. They could also be the result of or differences between the local average treatment effect (LATE) I estimate – that of being placed in (or not being placed in) special education as a result of AYP incentives – and the parameters estimated in prior work.

To investigate the first possibility, I replicate the main analysis from Hanushek, Kain, and Rivkin (2002) on my sample of students in North Carolina in the NCLB era. This specification uses student fixed effects to control for unobserved differences between those who receive special education and those who do not, and measures the effect of special education placement on gain scores.¹⁵ Here the change in score for student i in grade g in school j in year t (ΔA_{igt}) is a function of the student’s special education status in that year (SE_{igt}), student characteristics (X_{igt}), school characteristics (D_{gt}), a student fixed effect (γ_i), a school fixed effect (δ_j), cohort by grade dummies (ω_{gt}), and an error term (ϵ_{igt}):

$$(6) \Delta A_{igt} = SE_{igt}\lambda + X_{igt}\beta + D_{gt}\theta + \gamma_i + \delta_j + \omega_{gt} + \epsilon_{igt}.$$

Student characteristics include free or reduced-price lunch eligibility and an indicator for whether the student changed schools that year; school characteristics include the percentage of students who were Black, the percentage Hispanic, and the percentage eligible

¹⁵ I do not have the power necessary to include student fixed effects in my IV analysis.

for free or reduced-price lunch. I estimate the model as written, and removing the student fixed effects but adding controls for student race and gender. The resulting estimates appear in Table A5. Using this student fixed effect specification, I find small but positive effects of special education on student achievement in math, about twice the size of those found by Hanushek, Kain, and Rivkin. However, there is substantial variation across diagnosis groups and student ability as defined by third-grade test scores. Students diagnosed with learning disabilities or other health impairments experienced especially large gains, while there is no significant effect for those with autism. My sample includes a smaller proportion of students with learning disabilities and larger proportions with other health impairments (the classification used for ADD and ADHD) and autism. Students who scored in lower achievement levels in third grade saw greater gains in special education than did those who started with higher test scores. Taken together, these suggest that the differences between my estimates and those in previous work are not primarily driven by differences in sample.

6.3 Mechanisms

Why would my estimates be so different from those found in previous research? My estimates reflect the local average treatment effect (LATE) of being placed in (or not being placed in) special education as a result of AYP incentives. Recalling Figures 3 and 4, this mostly takes the form of students who had previously received a failing score either leaving special education or never entering it in the first place. Previous research has focused on the average treatment effect (ATE) for students who move in and out of special education or the LATE for students who were placed in special education because they were low-performing and attended schools that faced accountability pressure under a pre-NCLB system.

It is possible that the marginal special education students in my setting are harmed by the stigma, low-expectations, or lower-achieving peer group in a way that the average

special education student is not. It could also be that schools were intentional in pushing low-achieving students out of special education, particularly targeting those for whom it was the worst fit. These students might be easier to discourage from special education, and any gains they experienced by not being in special education could have the added benefit of helping the school achieve AYP in other subgroups. It could also be that when schools were forced to serve these students' needs outside of special education they turned to alternatives that were even better, or otherwise changed how they supported student learning. I explore several possibilities below. I use Equation (4) when considering mechanisms operating through school reactions, as I want to capture any potential channels through which incentives alter test scores. When considering student reactions to being placed in special education, I estimate Equation (5).

First, I consider whether schools substituted other supports for special education. Schools could be particularly likely to do so when trying to discourage special education for certain students, so might need to make a case that there is scope to meet the student's needs outside of special education. Supports might include extra time or attention, which is not observable in the data, or grade retention, which is. Students who are held back have another chance to master that grade's content. Recent research suggests that being held back in third grade improves the performance of students who struggle in reading, by 23% of a standard deviation in reading and 30% of a standard deviation in math (Schwerdt, West, & Winters, 2017). Students who are held back are also able to take an easier test – say the 4th grade test rather than the 5th grade test they would have taken if not held back – on which they are compared with younger students. A student taking the 4th grade test rather than the 5th grade test would be expected to perform roughly a half standard deviation better (North Carolina Public Schools, 2009). Taking these two effects together, a retained student

would be expected to receive test scores that were roughly 75 percent of a standard deviation better than if they had not been retained.

Table 7 displays the effect of accountability incentives on the likelihood that a student is promoted. I find no evidence that schools respond to selection incentives by changing promotion behavior, although schools are more likely to hold back students who scored in level 3 in reading when the school would benefit from their passing.

Table 7. Effect of Accountability Incentives on Being Promoted to the Next Grade

	(1)	(2)	(3)	(4)
Reading Level 2*Incentive	-0.000228 (0.00155)		0.000139 (0.00149)	
Math Level 2*Incentive	0.000468 (0.00204)		0.00276 (0.00166)	
Reading Level 3*Incentive	-0.00374* (0.00173)		-0.00345* (0.00170)	
Math Level 3*Incentive	-0.000745 (0.00128)		0.00276 (0.00150)	
Reading Level 2*Distance*Incentive		-0.000462 (0.00165)		-0.000183 (0.00159)
Math Level 2*Distance*Incentive		0.00125 (0.00227)		0.00257 (0.00188)
Reading Level 3*Distance*Incentive		-0.00437* (0.00204)		-0.00446* (0.00200)
Math Level 3*Distance*Incentive		-0.000961 (0.00154)		0.00227 (0.00192)
Reading Prior Pass*SWD Incentive	-0.00231 (0.00282)	-0.00255 (0.00286)		
Math Prior Pass*SWD Incentive	0.00315 (0.00488)	0.00383 (0.00505)		
Reading Prior Score*SWD Incentive			0.00173 (0.00196)	0.00179 (0.00195)
Math Prior Score*SWD Incentive			-0.000277 (0.00307)	-0.000177 (0.00303)
N	1199720	1199720	1199720	1199720

Notes: This table displays results from 4 linear probability models following Equation (4), one in each column, in which being promoted to the next grade is the dependent variable. Each model includes demographic controls, year-by-grade fixed effects, school-level fixed effects, and the main effects of school incentives and prior student performance. Standard errors in parentheses are clustered at the school level. * denotes significance at the 0.05 level, ** at the 0.01 level.

Table 8. Effect of Accountability Incentives on Changing Schools

	(1)	(2)	(3)	(4)
Reading Level 2*Incentive	-0.000313 (0.00280)		-0.000997 (0.00280)	
Math Level 2*Incentive	-0.00629* (0.00291)		-0.00465 (0.00284)	
Reading Level 3*Incentive	0.00302 (0.00360)		0.00115 (0.00351)	
Math Level 3*Incentive	0.00424 (0.00285)		0.00386 (0.00205)	
Reading Level 2*Distance*Incentive		-0.000902 (0.00305)		-0.00159 (0.00308)
Math Level 2*Distance*Incentive		-0.00793* (0.00319)		-0.00556 (0.00327)
Reading Level 3*Distance*Incentive		0.00235 (0.00402)		0.000530 (0.00389)
Math Level 3*Distance*Incentive		0.00476 (0.00368)		0.00451 (0.00262)
Reading Prior Pass*SWD Incentive	-0.00146 (0.00414)	-0.00107 (0.00417)		
Math Prior Pass*SWD Incentive	-0.00372 (0.00915)	-0.00453 (0.00929)		
Reading Prior Score*SWD Incentive			0.00173 (0.00196)	0.00179 (0.00195)
Math Prior Score*SWD Incentive			-0.000277 (0.00307)	-0.000177 (0.00303)
N	1199737	1199737	1199737	1199737

Notes: This table displays results from 4 linear probability models following Equation (4), one in each column, in which being in a new school in the current year is the dependent variable. Each model includes demographic controls, year-by-grade fixed effects, school-level fixed effects, and the main effects of school incentives and prior student performance. Standard errors in parentheses are clustered at the school level. * denotes significance at the 0.05 level, ** at the 0.01 level.

Schools could also attempt to change their accountability populations by encouraging or discouraging student movement in and out of schools. This could affect achievement either if students systematically move into better (or worse) schools or because changing schools is generally disruptive. If schools pushed students out (or held on to them) anytime a student was placed in special education this would be part of the policy effect of being in special education; if it only occurred in the presence of NCLB incentives it would be

part of the policy effect of being in special education under NCLB. I examine whether previously passing students are more likely to be in a new school when their school faces an incentive to alter the SWD population, with results appearing in Table 8. I find no evidence that schools respond on this margin when trying to improve the performance of the SWD subgroup, although students who previously scored in level 2 in math are more likely to be in a new school when their school would benefit from their passing.

It is also possible that students respond to their placement by changing their level of effort. This could be a reaction to stigma or low expectations, which might be particularly marked for students whose placements are altered by incentives. While many aspects of effort are difficult to observe, I use information on absences to determine whether special education affects one fundamental aspect of effort – attendance. This is both an indicator of overall effort and an input in itself – previous research suggests that each additional absence lowers math achievement scores by 0.05 of a standard deviation (Goodman, 2014).

The reading incentives form strong instruments in this sample, but the math incentives do not, as shown in Table 9. However, their pattern of signs and significance is similar to that using the reading incentives as instruments. I find no evidence that special education increases overall absences or excused absences, but I do find evidence that it increases unexcused absences and instances of being tardy. Using Goodman’s estimates, a 6 percentage point increase in absences would be expected to lower math achievement by 0.5 standard deviations.

Table 9. Effect of Special Education on Attendance

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel 1. Percent of Days Absent								
	-0.0138 (0.0352)	3.023 (0.0355)	-0.0115 (0.0344)	-0.0198 (0.0338)	0.140 (0.103)	0.0832 (0.0651)	0.146 (0.106)	0.0882 (0.0659)
Panel 2. Percent of Days Excused Absence								
	-0.00404 (.0107)	-0.0109 (0.0115)	-0.00448 (0.0106)	-0.0109 (0.0111)	0.0462 (0.0300)	0.0224 (0.0173)	0.0417 (0.0281)	0.0209 (0.0174)
Panel 3. Percent of Days Unexcused Absence								
	0.0203 (0.0118)	0.0286* (0.0135)	0.0197 (0.0117)	0.0267* (0.0128)	0.0903 (0.0472)	0.0619* (0.0271)	0.0859 (0.0454)	0.0591* (0.0271)
Panel 4. Times Tardy per Day Enrolled								
	0.0190* (0.00926)	0.0238* (0.0115)	0.0191* (0.00940)	0.0229* (0.0111)	0.0157 (0.0311)	0.0479 (0.0266)	0.0493 (0.0328)	0.0475 (0.0269)
F-statistic	19.34	18.04	19.56	19.22	4.89	9.881	4.927	9.606
N	928511	928511	928511	928511	928511	928511	928511	928511
Instruments:								
	Reading	Reading	Reading	Reading	Math	Math	Math	Math
Prior Pass * SWD Incentive	Y		Y		Y		Y	
Prior Score * SWD Incentive		Y		Y		Y		Y
Controls:								
Levels * Incentive	Y	Y			Y	Y		
Levels * Distance*Incentive			Y	Y			Y	Y

Notes: This table displays results from 24 linear IV models following Equation (5), one in each column and panel, in which being in special education with a malleable diagnosis is instrumented by selection incentives as noted. All specifications include demographic controls, year-by-grade fixed effects, and school-level fixed effects, and I control for hypothesis 1 incentives as noted. Standard errors in parentheses are clustered at the school level. Kleibergen-Papp F-statistics from the first stage are reported. * denotes significance at the 0.05 level, ** at the 0.01 level.

In sum, I do not find evidence that observable changes in schools' other investments in students drive the negative effects of special education on achievement for marginal students. However, it appears that students react to being placed in special education in ways that have negative implications for achievement. Prior work has shown that students in special education have worse attendance and report lower engagement with school and peers (Bear, Clever, and Proctor, 1991, Lackaye and Margalit 2006, Stiefel et al. 2017). My results suggest that these differences are at least in part causal rather than purely correlational, and highlight the need for a better understanding of how to mitigate the negative consequences of special education placement.

7. Conclusion

I examine school responses to AYP incentives to classify particular students as disabled in order to either target resources to students close to the proficiency threshold or to change the composition of the students with disabilities (SWD) subgroup. I use variation across schools in their past performance in the subjects and subgroups relevant to AYP, and across students in their prior scores and subgroup membership in order to isolate school responses.

I find evidence that schools discourage special education classification for students who have previously failed their reading or math test when the school benefits from improving the passing rate for the students with disabilities subgroup. I also find evidence that schools use special education to target resources to students near the passing threshold in reading when the school would benefit from their passing. However, students who just passed in math are unaffected, and those who just failed are less likely to be in special education when their school would benefit from their passing. This likely reflects two factors. First, state accountability rewards schools for the percentage of students passing, without a

focus on subgroups or a single high-stakes threshold, so that schools already attempt to improve the scores of almost- and just-passing students regardless of AYP incentives. Second, North Carolina's formula for funding special education incentivizes schools to limit the size of their special education population, so schools may discourage placement for some in order to "make space" for others.

While it is important to understand how schools have responded to policy incentives, it is not clear what those responses mean for students. Either over classification or under classification is at best an inefficiency and at worst an impediment to student learning and development. Without knowing the underlying need for special education services, it is unclear which prevails. It also is possible for the wrong students to be targeted even if neither over classification nor under classification occurs. I find that, for students whose placement is driven by their schools' incentives to alter the SWD population to be higher performing in math, special education is harmful to math achievement. Effects on reading achievement are consistently negative for this group but not significant. For students whose placement is influenced by the school's incentive to improve the performance of the SWD population in reading, there is no significant effect on reading or math scores.

This raises the question of why special education has different effects on these two groups of students. Differences across groups could be driven by differences in either the beneficial or detrimental effects of special education placement, or both. One possibility is that services and supports unique to special education have more scope to improve the performance of students who are low-performing in reading than in math. This seems plausible, as reading performance is generally harder for schools to alter, and the alterations in the NCEXTEND2 might be especially valuable to struggling readers. Another possibility is that schools are better at discouraging placement for students who are low-performing in

math but would not be well-served by special education than they are at discouraging placement for similar students who are low-performing in reading. This could either be because such students are more difficult to identify or because schools have less discretion in their placement.

School reactions to selection incentives mostly take the form of discouraging previously low-performing students from entering or remaining in special education. Thus, my results suggest that schools faced with accountability pressure are rationally using special education placement to serve their own goals, with benefits to some of the students affected. While schools that do not face accountability pressure might also benefit from discouraging special education placement for some students, it may be costly to do so. This could be because providing alternative supports is expensive and not defrayed by additional state funding, because identifying who would do better without special education is difficult, or because there are strong pressures to place low-performing students into special education.

My main estimates do not reflect the average benefit of special education for all students who receive it, but rather marginal special education students. These are the students whose placement can reasonably be altered by the action of stakeholders or plausible changes to identification and classification procedures. As such, their experiences are the ones relevant for setting accountability policy. Importantly, while the previous literature supports a policy of providing as much special education as budgets allowed, my finding suggests that placing a student into special education can in fact be harmful to achievement. Thus, it is crucial to target special education services to students who will benefit from them.

These are to my knowledge the first estimates of how schools responded to AYP incentives to alter which students received special education. In comparison with earlier accountability regimes, NCLB appears to have eliminated one method of gaming the system - removing low-achieving students from the accountability population - and replaced it with others – targeting special education to students near the passing threshold in reading and manipulating subgroup composition. Although NCLB is no longer in force, current accountability policies impose similar incentives on schools with a continued emphasis on the percentage of students meeting targets and the use of a students with disabilities subgroup. These facts – that schools can and will manipulate special education placement in the face of NCLB-style incentives and that some students can be hurt by special education placement- are important for policymakers and stakeholders to consider as accountability and special education policies continue to evolve in the future.

REFERENCES

- Angrist, J. & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.
- Bear, G.G., Clever, A., & Proctor, A. (1991). Self-perceptions of non-handicapped children and children with learning disabilities in integrated classrooms. *The Journal of Special Education*, 24(4), 409-426.
- Bokhari, F. & Schneider, H. (2011). School accountability laws and the consumption of psychostimulants. *Journal of Health Economics*, 30(2), 355-72.
- Byrd-Blake, M., Afolayan, M. O., Hunt, J. W., Fabunmi, M., Pryor, B. W., & Leander, R. (2010). Morale of teachers in high poverty schools: A post-NCLB mixed methods analysis. *Education and Urban Society*, 42, 450–472.
- Chakrabarti, R. (2013). Accountability with voucher threats, responses, and the test-taking population: Regression discontinuity evidence from Florida. Federal Reserve Bank of New York Working Paper.
- Cohen, J. (2007). Causes and consequences of special education placement: Evidence from Chicago Public Schools. Brookings Institution Working Paper.
- Cullen, J. (2003). The impact of fiscal incentives on student disability rates. *Journal of Public Economics*, 87(7-8), 1557-1589.
- Cullen, J. & Reback, R. (2006). Tinkering toward accolades: school gaming under a performance accountability system. In T. J. Gronberg and D. W. Jansen (Eds.), *Improving School Accountability (Advances in Applied Microeconomics, Volume 14)* (pp. 1 - 34). Emerald Group Publishing Limited.
- Dee, T. S., & Jacob, B. A. (2010). The impact of No Child Left Behind on students, teachers, and schools. *Brookings Papers on Economic Activity*, 149–194.
- Federal Education Budget Project (2014). Background and analysis: Individuals with Disabilities Education Act – Cost impact on local school districts. Retrieved from <http://febp.newamerica.net/background-analysis/individuals-disabilities-education-act-cost-impact-local-school-districts>
- Figlio, D. (2006). Testing, crime, and punishment. *Journal of Public Economics*, 90 (2006), 837-851.

- Figlio, D. & Getzler, L. (2002). Accountability, ability, and disability: Gaming the system. In T. J. Gronberg and D. W. Jansen (Eds.) *Improving School Accountability (Advances in Applied Microeconomics, Volume 14)* (pp. 35 – 49). Emerald Group Publishing Limited.
- Figlio, D. & Winicki, J. (2008). Food for thought: the effect of school accountability plans on school nutrition. *Journal of Public Economics, 89*(2–3), 381-394.
- Hanushek, E., Kain, J., & Rivkin, S. (2002). Inferring program effects for special populations: Does special education raise achievement for students with disabilities? *The Review of Economics and Statistics, 84*(4), 584-599.
- Hanushek, E. & Raymond, M. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297-327.
- Jacob, B. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*(5-6), 761-796.
- Kwak, S. (2010). The impact of intergovernmental incentives on student disability rates. *Public Finance Review, 38*(1), 41-73.
- Griffith, G., & Scharmann, L. (2008). Initial impacts of No Child Left Behind on elementary science education. *Journal of Elementary Science Education, 20*, 35–48.
- Lackaye, T. D. & Margalit, M. (2006). Comparisons of achievement, effort, and self-perceptions among students with learning disabilities and their peers from different achievement levels. *Journal of Learning Disabilities, 39*(5). 432-446.
- Ladd, H. F. & Lauen, D. L. (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management, 29*, 426–450.
- Mahitivanichcha, K. & Parrish, T. (2005). The implications of fiscal incentives on identification rates and placement in special education: Formulas for influencing best practice. *Journal of Education Finance, 31*(1), 1-22.
- Morrill, M.S. (2016). Special education financing and ADHD medications: A bitter pill to swallow. Working paper. Retrieved from https://sites.google.com/a/ncsu.edu/msmorrill/files/Morrill_BitterPill.pdf?attredirects=0
- Neal, D. & Schanzenbach, D. (2010). Left behind by design: proficiency counts and test-based accountability. *The Review of Economics and Statistics, 92*(2), 263-283.
- North Carolina Public Schools (2009). North Carolina Extend2 end-of-grade reading comprehension tests technical report. Edition 2. Retrieved from <http://www.ncpublicschools.org/docs/accountability/testing/reports/extend2readingtechmanual.pdf>

U.S. Department of Education, Office of Special Education and Rehabilitative Services (2010). A guide to the Individualized Education Program. Jessup, MD: ED Pubs.

U.S. Department of Health and Human Services, Office for Civil Rights (2006). Your rights under Section 504 of the Rehabilitation Act. Retrieved from <https://archive.hhs.gov/ocr/504.pdf>

Reback, R. (2008). Teaching to the rating: school accountability and the distribution of student achievement. *Journal of Public Economics*, 92(5-6), 352-353.

Schwerdt, G., West, M. R., Winters, M.A. (2017). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. *Journal of Public Economics*, 152, 154-169.

Setren, E. (2016). Special education and English language learner students in Boston charter schools: Impact and classification. Working Paper. Retrieved from <http://economics.mit.edu/grad/esetren/research>

Stiefel, L., Shiferaw, M., Schwartz, A. E., and Gottfried, M. (2017). Is special education improving? Evidence on segregation, outcomes, and spending from New York City. IESP Working Paper No. 02-01.

U.S. Department of Education, National Center for Education Statistics. (2013). Digest of Education Statistics, 2012 (NCES 2014-015), Table 48.

U.S. Department of Education, National Center for Education Statistics. (2016). Digest of Education Statistics, 2014 (NCES 2016-006), Table 236.60.

Winters, M. & Greene, J. (2011). Public school response to special education vouchers: The impact of Florida's McKay scholarship program on disability diagnosis and student achievement in public schools. *Educational Evaluation and Policy Analysis*, 33(2), 138-158.

CHAPTER 2

EFFECTS OF DI WAIT TIME ON HEALTH AND FINANCIAL WELL-BEING

1. Introduction

In order to qualify for Social Security Disability Insurance (DI) benefits, applicants must demonstrate that they are unable “to engage in any substantial gainful activity (SGA) by reason of a medically determinable physical or mental impairment(s) which can be expected to result in death or which has lasted or can be expected to last for a continuous period of not less than 12 months” (Social Security Act, 1965). Engaging in SGA is defined as earning more than a set amount per month, \$1,180 for non-blind individuals in 2018. In 2008, the average DI applicant waited about four months for an *initial* determination. However, because those whose claims are denied can appeal that decision at multiple levels, total waiting time is often much longer. On average, in 2008 applicants waited over a year for a final decision, with around 10 percent of applicants waiting three years or longer (Autor, Maestas, Mullen, & Strand, 2015).

A long wait can mean an extended period out of the labor force as well as delayed benefits and insurance coverage. Applicants who are engaging in SGA are not eligible for benefits, and employment while waiting for a decision can be used as evidence that the applicant is able to work. Beneficiaries are eligible for DI benefits five months after onset or after a favorable decision. Although retrospective payments are made for months when the beneficiary was waiting for a decision, applicants do not know if they will be awarded benefits, or when that award will occur, which may make consumption smoothing difficult

for many.¹⁶ Applicants who are still waiting for a decision 29 months after disability onset would also experience a delay in Medicare coverage. For the 16% of DI applicants who apply to Supplemental Security Income (SSI) at the same time, even the average four month wait for an initial decision means delayed financial support and insurance coverage, as they would receive SSI benefits Medicaid eligibility essentially immediately upon acceptance. For those who are ultimately denied there is no compensation for time spent waiting.

Time out of the labor force, delays in receiving benefits and insurance coverage, and uncertainty can all have implications for applicants' well-being by affecting their daily activities, income, and health care access. Staying out of the labor force means time for job skills and labor force attachment to decay. It could lower well-being through a drop in income, which limits both consumption and applicants ability to spend money to improve their health, and loss of employer-based health insurance. However, it could allow applicants to invest more time in their health, which might be particularly important when recovering from an illness or adjusting to new impairments. There is some evidence that receiving DI for a short period of time increases earnings for those who are subsequently removed from the program, which might be explained by this opportunity to invest time in health at a crucial moment (Moore, 2015). Delayed benefits mean lower income, at least temporarily, while delayed Medicare or Medicaid coverage could harm both health care access and financial well-being, as applicants are forced to either forgo insurance coverage or pay for more expensive, and perhaps less comprehensive, coverage (Gross & Notowidigdo, 2011).¹⁷

¹⁶ In a survey of 2008 DI awardees, 80% reported that waiting for benefits had affected their finances. Two thirds of these reported relying on assistance from friends, family, and charity, while 40% took on debt (SSA 2009).

¹⁷ Analyses using a similar data set to that used here found that almost a quarter of beneficiaries were without insurance coverage in the year before application, 37 percent had insurance from their own

Research suggests that health insurance improves access to health care, as well as some components of health (Currie & Gruber, 1996, Michalopoulos et al., 2011, Finkelstein et al., 2012, Baicker et al., 2013). This may be especially important for DI applicants, who by definition have a serious health care condition.

A well-established literature has estimated the effect of DI on applicants' and beneficiaries' labor force participation and earnings (Bound, 1989, Chen & van der Klaauw, 2008, Maestas Mullen & Strand, 2013, French & Song, 2014, Gelber, Moore & Strand, 2017). Less is known about the effects of DI on other outcomes. In a recent working paper Gelber, Moore & Strand (2018) find that larger benefit amounts decrease mortality among beneficiaries in the first five years after allowance. As discussed above, research also suggests that, for those who lose eligibility for benefits, having received benefits for about three years can increase earnings compared with those who have been on for a very short time (Moore, 2015). An analysis of changes to the Dutch DI system suggests that changes to income and work can have implications for the health of DI beneficiaries. However, the Dutch systems of DI and social support are quite different from those in the US, so this work provides limited insights into the US system (Garcia-Gomez & Gielen, 2014).

Wait time appears to play a substantial role in the effect of DI - taking wait time into account increases the effect of DI on employment by about 50%, suggesting that the previous consensus understated the impact of the program substantially (Autor et al., 2015). Long waits force applicants to change their behaviors in order to fund consumption. Coe, Lindner, Wong, & Wu (2013) investigate the coping strategies that applicants use, and find that longer waiting times increase SNAP usage while decreasing the use of unemployment

employer, and 33 percent had insurance through a spouse's employer (Livermore, Stapleton, and Claypool, 2009).

insurance and the likelihood of changing addresses. Despite these indications that waiting time is important to well-being, to my knowledge no other research has addressed the effect of DI wait time in a setting that allows for causal inference.

I address this gap in the literature using the restricted 1997-2005 National Health Interview Survey (NHIS) linked to two SSA administrative files – the Master Beneficiary Record (MBR) and 831 File – through 2007 to examine the effects of waiting time on health, health care access, and financial well-being. This linked data combine the relative accuracy and programmatic detail of administrative records with a rich description of well-being available only in survey data.

Individual wait time depends in part on the characteristics and choices of applicants. To simply compare the outcomes of those with different wait times would conflate the effect of time waiting for a decision with these factors. Instead, I use information on the number of pending applications and number of decisions made to construct expected wait times by state and month of application, and use these as instruments for individual wait time. This allows me to isolate the variation in wait time that is caused by factors beyond the individual's control, such as differences in processing speed or backlogs from earlier applications.

I find evidence that wait time increases the likelihood of currently receiving benefits at the time of survey and decreases that of having had benefits terminated. This is broadly consistent with previous findings that wait time decreases employment and earnings, and would be expected if a longer wait makes the return to work more difficult. I find that wait time decreases the likelihood of seeking a reconsideration. I also find evidence that wait time increases the number of conditions causing activity limitations. This may point to one of the ways in which wait times impede return to work, but also demonstrates that wait time has

implications for beneficiaries' well-being that are not confined to workforce outcomes. I do not find significant effects of wait time on other outcomes, although point estimates suggest that wait time increases BMI, has no effect on mental health, and increases poverty. Ultimately, I am limited in my ability to identify the effects of wait time by sample size.

My results provide evidence that the effects of DI, and of wait time in particular, are not limited to workforce outcomes. Researchers have so far focused on these work and earnings because they are convenient to study. However, while employment and earnings are important, they are insufficient to characterize well-being or to measure the impact or value of DI.

The rest of this paper proceeds as follows. Section 2 describes the data used and samples constructed. In Section 4 I describe the method used, and in Section 5 the results. Section 6 concludes.

2. Data and Sample

2.1 Data Sources

I use linked data that includes information from the 1997-2005 National Health Interview Survey (NHIS), the NHIS restricted access mortality file through 2011, and two SSA administrative files – the Master Beneficiary Record (MBR) and 831 File – through 2007. The NHIS is a cross-sectional household survey that covers the civilian non-institutionalized population of the United States (National Center for Health Statistics, 2006). Respondents are asked to provide their social security numbers and consent to have their survey responses linked to other data. From the survey I draw information on demographics, state of residence, and a host of indicators of health, health care access, and financial well-being. Because many of the indicators of health care access measure similar concepts, I create an index of health care access. To create the index I count the number of questions about health

care access with non-missing data for each respondent, as well as the number of questions on which the respondent reported a barrier to access¹⁸. I then divide latter number by the former number, resulting in an index runs from 0 to 1.

The MBR and 831 files are used by the Social Security Administration for program operations. The MBR includes information, in many cases monthly, on all individuals who apply for DI or for retirement benefits from SSA. The 831 file is focused on the determination process for those who apply to either DI or SSI. My extract of these files includes only those individuals who are matched to the NHIS data and appear in both the MBR and the 831 file. Crucially, the files record the dates on which applications were submitted, and decisions made. They also include information on primary disabling conditions, month-by-month program status, and other programmatic details.

2.2 Sample

The linked data provides information for all NHIS respondents who provided SSNs and consented to be linked, for whom a successful link was performed, and who applied to DI between 1988 and 2007. From this file I construct a sample that includes those who applied before interview. For my analyses I use an instrument defined using summary statistics on the number of applications received and processed in each state and month. This information is only available beginning in October 2000, so my sample is further restricted to those with an application on or after that month. For individuals with more than one application I find the most recent initial application at time of survey and consider this as

¹⁸ These questions include: whether the respondent has seen a dentist in the past 12 months; whether prescription medicine, mental health care, dental care, or medical care were needed but could not be afforded; whether care was delayed due to cost; whether the respondent has a usual place for care; whether the respondent has no form of health insurance; and whether the household paid \$500 or more out of pocket in the past year.

their application of interest¹⁹. I identify the initial decision as the earliest decision associated with that initial application date, and note all decisions associated with that application date recorded in the 831 file.²⁰ I drop the handful of applications considered under the Quick Disability Determination program and other special expedited processes. I also drop those for whom the wait time for the initial decision cannot be determined due to missing or inconsistent information, those with negative wait times, and those with wait times in the top 1%. These final two restrictions come from an assumption that implausible wait times are more likely to be reflective of data errors than truth, and that very extreme waits are unlikely to be driven by variation in expected wait times.

The main sample includes 2,155 individuals who applied for benefits from October 2000 to December 2005. On average, they were 47.5 years old at survey and 46 at application, as shown in Table 10. Over a third of applicants had a musculoskeletal primary disabling condition and 22 percent had a mental health condition as their primary diagnosis. On average, they faced a 102 day initial wait time, about 3 ½ months, although this average camouflages the long right tail in wait times. This is more apparent in Figure 5, which depicts the distribution of initial wait times, top coded at 300 days.

¹⁹ Applications begin with an initial application and initial decision. Individuals whose claims were rejected could continue to pursue their applications through several levels of appeal. They can also resubmit their claim after a period of time.

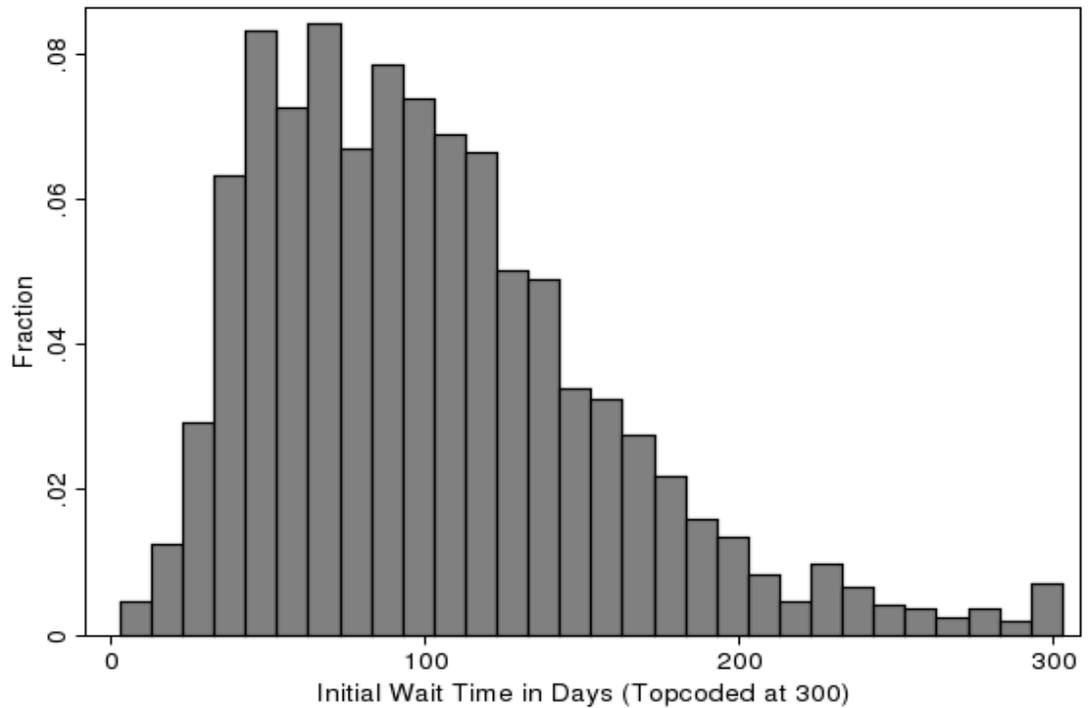
²⁰ Unfortunately, the 831 file does not record detailed information on most appeals beyond the reconsideration step. The MBR contains some information on these decisions, but my extract of the file does not include them in the format needed to accurately trace their path.

Table 10. Descriptive Statistics – Demographics and Wait Time

	N	Mean	SD
Age at Survey	2,143	47.455	11.907
Age at Application	2,143	45.914	11.836
Male	2,148	0.476	0.500
Race/Ethnicity			
White	2,155	0.477	0.500
African American	2,155	0.177	0.382
Hispanic	2,155	0.162	0.369
Other	2,155	0.183	0.387
Education			
< HS	2,155	0.220	0.414
HS or GED	2,155	0.325	0.468
> HS	2,155	0.237	0.425
Unknown	2,155	0.219	0.413
Marital Status			
Married/Partnered	2,155	0.578	0.494
Widowed	2,155	0.049	0.216
Divorced/Separated	2,155	0.208	0.406
Never Married	2,155	0.162	0.368
Unknown	2,155	0.002	0.048
Primary Disabling Condition			
Musculoskeletal	2,155	0.345	0.476
Senses and Speech	2,155	0.026	0.159
Respiratory	2,155	0.045	0.208
Cardiovascular	2,155	0.095	0.293
Digestive	2,155	0.022	0.148
Genito-urinary System	2,155	0.013	0.113
Endocrine	2,155	0.050	0.217
Neurological	2,155	0.075	0.263
Mental	2,155	0.222	0.416
Neoplastic	2,155	0.032	0.175
Immune	2,155	0.017	0.130
Other	2,155	0.046	0.210
Initial Wait	2,155	102.357	56.627
Reconsideration	2,155	0.361	0.481
Initial + Reconsideration Wait	2,155	146.339	102.442
Reports Applying for SSDI	2,134	0.724	0.447
Reports Applying for SSI	2,131	0.363	0.481
Applied for SSI	2,155	0.550	0.498
Correctly Reports SSI application	2,131	0.618	0.486

Notes: Table presents descriptive statistics on analytic sample constructed from the 1997-2005 NHIS linked to 1989-2005 SSA administrative records.

Figure 5. Distribution of Initial Wait Times



Notes: Figure displays the fraction of applicants who fall into 10-day initial wait bins, with the top bin containing those whose waits were 300 days or longer.

Linking survey data to administrative records allows me to evaluate how accurately survey respondents reported their application behavior. Although all sample members had applied for DI at the time of survey, less than 75% reported doing so. A little more than half had applied for SSI at the time of survey, but about a third reported doing so. Only around 61 percent correctly reported their SSI application status. These inaccuracies are part of the reason administrative data is crucial to analyzing beneficiary experiences.

Table 11. Descriptive Statistics – Program Status, Health, and Well-being

	N	Mean	SD
Benefit Status at Survey			
Current	2,155	0.591	0.492
Suspended	2,155	0.009	0.094
Terminated	2,155	0.064	0.244
Never Benefits	2,155	0.337	0.473
# of Activity Limitations	2,155	1.698	1.914
# of Functional Limitations	2,155	1.048	1.689
Good or Better Health	2,151	0.397	0.489
>20 bed days	1,073	0.333	0.471
Number MH symptoms	1,130	1.325	1.924
Any MH symptoms	1,130	0.443	0.497
BMI	1,139	36.337	20.134
Dentist Past 12 Months	1,124	0.459	0.499
Unable to afford needed:			
Prescription medicine	1,129	0.324	0.468
Mental health care	1,127	0.147	0.355
Dental Care	1,128	0.318	0.466
Needed Medical Care	2,151	0.313	0.464
Delayed Care - Financial	2,152	0.346	0.476
Has a Usual Place for Care	1,116	0.899	0.302
No Health Insurance	2,143	0.247	0.431
Working Last Week	2,144	0.171	0.376
SNAP	2,142	0.208	0.406
>\$500 out of pocket	2,110	0.580	0.494
< 100% FPL	1,735	0.285	0.451
Dead by 12/31/2011	2,155	0.146	0.353

Notes: Table presents descriptive statistics on analytic sample constructed from the 1997-2005 NHIS linked to 1989-2005 SSA administrative records

Table 11 details sample sizes, means, and standard deviations for program status and the various measures of health, health care access, and financial well-being. At the time of survey about 60 percent were receiving DI benefits, 1 percent had benefits suspended, 6 percent had benefits terminated, and the remainder had never had benefits. About 40 percent reported being in good or better health. About 45 percent reported that they experienced at least one symptom of depression always or often, suggesting that many more than the 22 percent of sample members who had a mental health condition as their primary diagnosis might benefit from mental health care. In the year prior to interview, 46 had seen a dentist, 32 percent had been unable to afford prescription medicine, 15 percent mental health care, and 32 percent dental care. Thirty one percent had had some family member forgo needed medical care because it was too expensive. Thirty five percent had delayed medical care due to cost. About a quarter did not have health insurance at the time of survey. Despite these barriers, only 10 percent did not have a usual place for care. Almost 30 percent had family income at or below the Federal Poverty Level, and about 15 percent had died by the end of 2011.

3. Method

Simply comparing the outcomes and characteristics of those with shorter and longer wait times would conflate the true effects of waiting for a decision and the circumstances that cause some to face longer waits than others. Wait times vary across individuals both for reasons associated with that individual's outcomes, such as impairment, job prospects, or choices to pursue initially denied applications, and those that are not, such as examiner speed, determination office staffing, and previous caseloads. In order to identify the causal effect of wait time on outcomes it is necessary to isolate the variation caused by the latter factors. To do so I use publicly available information on the number of decisions and pending

applications for each state for each month from October 2000 to the present to construct the expected wait times for an initial decision that prevailed when and where the application was made, and use these as instruments.

$$(1) \text{ExpWait}_{sym} = \frac{\text{Pending}_{sym-1}}{\text{Decisions}_{sym}}$$

The expected wait (ExpWait_{sym}) for state s year y and month m is the number of pending applications for that state at the end of month $m-1$ divided by the number of decisions made in month m . This reflects the number of months it would take the Disability Determination Service to process an application submitted at the beginning of month m , assuming applications are considered in the order they are received and decisions are made at the rate that prevails in that month. The distribution of expected wait times is depicted in Figure A1.

$$(2) \text{Wait}_{isym} = \beta_1 X_{isym} + \beta_2 \text{ExpWait}_{sym} + \gamma_s + \gamma_y + \gamma_m + \text{Unemp}_{sym} + \varepsilon_{isym}$$

$$(3) \text{Outcome}_{isym} = \beta_1 X_{isym} + \beta_2 \widehat{\text{Wait}}_{isym} + \gamma_s + \gamma_y + \gamma_m + \text{Unemp}_{sym} + \varepsilon_{isym}$$

I use this instrument in a standard linear IV framework. In the first stage, wait time for individual i who applies in state s , year y , and month m (Wait_{isym}) is estimated based on individual demographics (X_{isym}), the instrument for expected wait (ExpWait_{sym}), state, year, and month fixed effects, the unemployment rate (Unemp_{sym}), and a random error. In the second stage, actual individual wait time is replaced by the estimated wait time produced by the first stage ($\widehat{\text{Wait}}_{isym}$). I report estimates with and without a control for the expected wait time from the previous month (ExpWait_{sym-1}). Standard errors are clustered at the state level.

My estimates reflect the causal effects of wait time if, conditional on controls, individuals who applied in state-months with different expected wait times would have the same outcomes were it not for the wait they face. In addition to demographics, (X_{isym}) I control for several other factors that might cause correlation between expected waits and individual outcomes. First, some states, months, and years have higher wait times in general, which may be correlated with unobserved differences in other characteristics. For example, applicants in December may be different from those in other months, and also face different wait times. I address this concern by including fixed effects for state, month, and year of application.

Second, applicants could know something about wait times and decide when to apply based on that information. This is unlikely to be a major factor for those who are not working just before application, but could be a consideration for disabled workers deciding to leave a job and pursue DI benefits. To address this issue, I make use of the fact that individuals do not know the wait they will face in advance, as summary data cannot be published until after the month has ended. Instead, individuals attempting to time their applications would have to rely on information on previous waits or previous backlogs to form their expectations. I include estimates that control for a 1-month lagged expected wait, reflecting what an applicant's best guess of their own wait time might be.

Third, applications generally increase when employment prospects are poor, lengthening wait times (Autor and Duggan, 2003). This is a problem both because the applicants who are induced to apply by poor economic conditions are probably different from other applicants and because the state of the economy can have independent effects on some of the outcomes I consider. I address this in two ways. First, I include a control for the unemployment rate in the state, year, and month of application to capture month-to-

month variation in economic trends. While unemployment does not fully capture applicants' job prospects it should be at least correlated with month-to-month changes in economic conditions. Second, to the extent that economic conditions evolve somewhat smoothly over time, the inclusion of lagged wait time should control for the effect of a similar economic situation, as it is reflected in wait time.

4. Results

4.1 First Stage

I begin by evaluating the strength of my instrument, with the results displayed in Table 12. Kleibergen-Papp F-statistics appear at the bottom of the table and are greater than 10 for the full sample. An additional month of expected wait causes 16.87 days of additional wait, as shown in column 1. After controlling for lagged wait, an additional month of expected wait causes 12.15 days of individual wait, reflecting serial correlation in waits.

Table 12. First Stage

	(5)	(6)
Instrument	16.87** (3.566)	12.15** (2.622)
Lagged Instrument		9.982** (2.575)
N	1,757	1,700
F-Stat	33.12	41.41
Lagged Wait Time	N	Y

Notes: Table reports results from 2 linear regressions of individual initial wait time on instruments. Each regression includes controls for age, age squared, sex, education, race/ethnicity, marital status, primary diagnosis, and the unemployment rate in the state and month of application. Standard errors are clustered at the state level. The Kleibergen-Paap F-statistic on the instrument is reported at the bottom of the table. + denotes significance at the 0.1 level, * at the 0.05 level, and ** at the 0.01 level.

Table 13. Effects of Wait Time on Program Status

	Benefit Status at Survey					
	Ever SSDI (1)	Current (2)	Suspended (3)	Terminated (4)	Never Benefits (5)	Any Re- consideration (6)
Panel 1 – No Lag						
Wait	-0.00027 (0.00073)	0.0013 (0.0010)	0.00023 (0.00026)	-0.0014* (0.00058)	-0.00014 (0.00078)	-0.0024* (0.0010)
F-statistic	33.12	33.12	33.12	33.12	33.12	33.12
N	2132	2132	2132	2132	2132	2132
Panel 2 – 1 Month Lag						
Wait	0.00026 (0.0011)	0.0026* (0.0012)	0.000050 (0.00038)	-0.0011 (0.00080)	-0.0015 (0.00095)	-0.0032* (0.0013)
F-statistic	41.41	41.41	41.41	41.41	41.41	41.41
N	2065	2065	2065	2065	2065	2065

Notes: Table reports the estimates of the effect of wait time on outcomes, from 12 IV regressions. The second panel controls for 1-month lagged wait time. Standard errors are clustered at the state level. Kleibergen-Paap F-statistic on the instrument is reported at the bottom of each panel. + denotes significance at the 0.1 level, * at the 0.05 level, and ** at the 0.01 level.

4.2 Effects of Wait Time

Table 13 displays IV estimates of the effect of wait time on program status. The estimates in Panel 2 include a control for the previous month's wait time, while those in Panel 1 do not. An additional day of wait time increases the likelihood of having benefits terminated at the time of survey by 0.14 percentage points, or about 2 percent. Most terminations occur due to death or aging out of the program. No one in my sample has died at the time of survey and the regression controls for age, so these effects must be driven by other terminations, such as those for work or medical recovery. Having benefits terminated for these reasons often takes time, so this could simply reflect the fact that those who face longer initial waits have received benefits for less time. It could also reflect changes to underlying factors, such as health or willingness and ability to work. It is worth noting that the effect of wait time on terminated status is not significant after controlling for lagged

wait, although the point estimate is reasonably close, and I find an increased likelihood of having current benefits at the time of survey. This leads me to believe that the loss of significance is result of a loss of power.

Longer wait times also decreased the likelihood of seeking a reconsideration by 0.24 percentage points when not controlling for lagged wait time, a 0.7 percent change. This is in contrast to the finding that wait time increased reconsiderations in Autor et al. (2015). The source of this difference is unclear, but it may result from differences in instrument. Their instrument identifies the effect of examiner speed, after controlling for average speed in the state, month and year of application. It seems possible that absolute wait time decreases reconsiderations, perhaps because applicants cannot afford to continue, but wait time relative to peers increases reconsiderations, because applicants interpret their long wait as a signal that theirs was a close decision that is likely to be overturned.

I do not find a significant effect on the likelihood of having ever received benefits at the time of survey, or of having benefits suspended. The point estimates are quite small and vary considerably across specifications. Greater precision would be needed to determine the effect of wait time on these outcomes. Estimates using survey weights appear in Table A.7 They are largely similar to those without weights, except that I find a marginally significant effect of wait time on ever having had benefits by the time of survey when controlling for lagged wait.

Table 14. Effects of Wait Time on Health and Well-Being

	Activity Limitations (1)	Functional Limitations (2)	Good or Better Health (3)	Bed Days (4)	BMI (5)	Mental Health Symptoms (6)	Health Care Access Index (7)	SNAP (8)	<= 100% FPL (9)	Working Last Week (10)
Panel 1 – No Lag										
Wait	0.0077* (0.0035)	0.0052 (0.0040)	-0.00076 (0.0018)	0.25 (0.69)	0.026 (0.020)	0.00013 (0.0016)	-0.0010 (0.00091)	-0.00033 (0.00091)	0.00061 (0.0012)	-0.00028 (0.00080)
F-statistic	33.12	33.12	32.40	10.40	8.286	10.33	7.576	30.45	29.58	31.29
N	2132	2132	2128	1131	969	1122	1079	2120	1718	2121
Panel 2 – 1 Month Lag										
Wait	0.014** (0.0037)	0.0049 (0.0060)	-0.0013 (0.0018)	-0.040 (1.32)	0.030 (0.030)	0.00066 (0.0026)	-0.00075 (0.0014)	-0.000034 (0.0015)	0.0019 (0.0013)	0.000036 (0.0010)
F-statistic	41.41	41.41	40.50	9.513	7.814	9.379	6.013	37.87	34.45	37.53
N	2065	2065	2061	1089	933	1080	1038	2053	1663	2055

Notes: Table reports the estimates of the effect of wait time on outcomes, from 20 IV regressions. The second panel controls for 1-month lagged wait time. Standard errors are clustered at the state level. Kleibergen-Paap F-statistic on the instrument is reported at the bottom of each panel. + denotes significance at the 0.1 level, * at the 0.05 level, and ** at the 0.01 level.

Table 14 presents estimates of the effect of wait time on health, health care access, and financial well-being, following the same structure as in the previous table. I find that an additional month of wait increases the number of conditions that cause activity limitations by about 0.2. The average respondent reports 1.7 such conditions, so this is a sizable increase, and controlling for 1-month lagged wait time doubles the estimated coefficient. I do not find significant effects of wait time on the number of conditions causing functional limitations, having good or better health, the number of bed days, BMI, or experiencing at least one depression symptom often or always. In some cases these point estimates would be meaningful if more precisely estimated. For example, more precise estimates would allow me to conclude that wait time has a substantial effect on the number of bed days, and no effect on the likelihood of reporting depression symptoms. Instead, I am unable to state with confidence that waiting has any effect on bed days, and cannot rule out changes to depression symptoms as large as 20 percent from an additional month of wait time.

The remainder of Table 14 displays estimates of the effect of wait time on the health care index, receiving SNAP in the previous year, having household income at or below 100% of the Federal Poverty Level, and working in the week previous to survey. None of these results are significantly different from zero. In most cases the point estimates suggest that wait time does not have an economically meaningful effect, but they are too imprecisely estimated to say this with confidence. The exception to this rule is poverty. Focusing on the results with a control for 1-month lagged wait time, the point estimate suggests that an additional month of waiting increases the likelihood of having a family income at or below FPL by about 6 percentage points, or 20 percent. Unfortunately, this estimate is very imprecise, so I am unable to rule out increases as large as 13 percentage points or decreases as large as 2 percentage points. Using survey weights does not substantively change these

results, as shown in Table A.8. I also find a similar pattern for each of the components of the health care access index, results for which appear in Table A.9.

It is likely that wait time matters more for some applicants than for others. For example, ultimately denied applicants never receive any compensation for their time waiting for a decision, while those who eventually receive benefits can receive back payments. Those with a spouse, especially a working spouse, may be better able to cope financially and have an easier time obtaining health insurance. This would be consistent with the finding by Coe et al. (2013) that applicants with employed spouses had longer wait times. Unfortunately, the size of my sample does not provide the power needed to investigate effects of wait time on most subgroups.

5. Conclusion

As of February 2018, over 10 million individuals received benefits from the DI program, totaling nearly \$11 billion in that month alone (SSA 2018). Little work has addressed the relationship between the DI program and outcomes other than employment, or the effects of the application process.

I use an instrumental variables strategy to estimate the effect of waiting for a decision on health, health care access, and other measures of well-being. I find that wait time decreases the likelihood of having had benefits terminated at the time of survey and increases the likelihood of currently receiving benefits at survey. I also find that a longer wait decreases the likelihood of asking for a reconsideration of a denied claim, and increases the number of conditions causing activity limitations. The sample I use is quite small, resulting in limited power. I find suggestive evidence that wait time affects some other outcomes, such as poverty and BMI, and does not affect others, such as mental health, but am unable to rule out alternate conclusions with confidence.

My analyses carry several additional limitations. Wait time exhibits a long right tail, but the instrument does not, making it unsuitable to investigate the effects of particularly long waits. I am also unable to address the effect on subsamples, as doing so results in very weak instruments. Finally, the effects captured here are relatively short-term. My sample applied for benefits no earlier than 2000, and were surveyed no later than 2005. If the effect of wait time changes over time, I am unable to comment on that here.

Despite these limitations, my findings highlight the fact that DI, and its wait time, affect more than work and earnings. Many applicants are not marginal workers. Even for those who are, the experience DI is not well summarized by decisions of whether to work and how much to earn. Focusing only on these outcomes makes it impossible to take account of the full value and cost of DI and other programs. It also does not provide answers to why programs have the labor market effects they do, which can be crucial to improving policy.

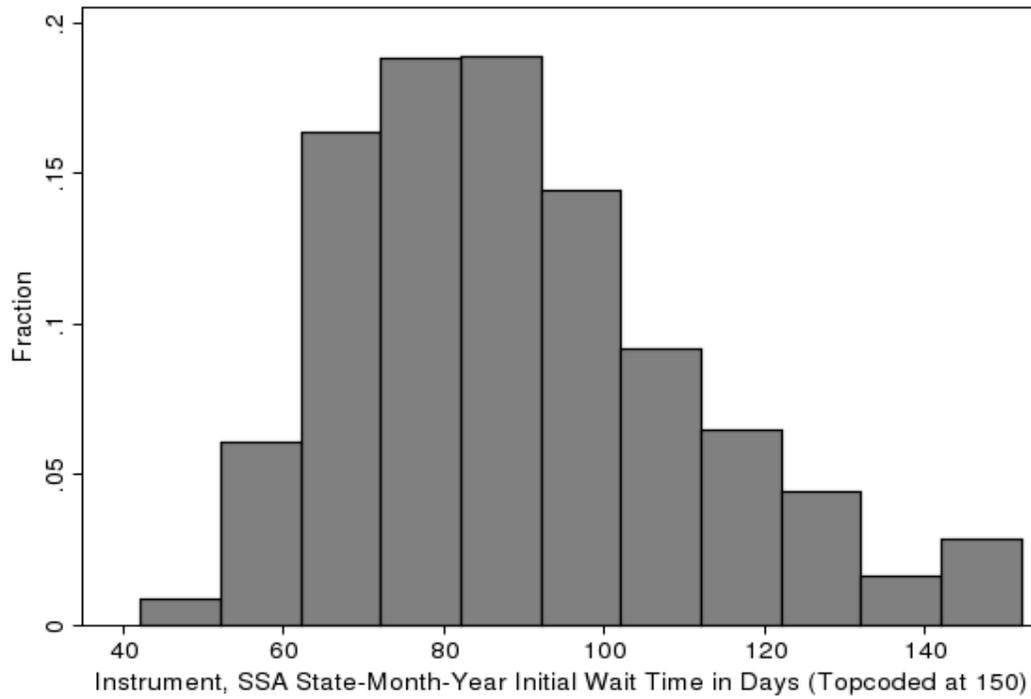
REFERENCES

- Autor, David, and Duggan, Mark (2003). The rise in the disability rolls and the decline in unemployment. *The Quarterly Journal of Economics*, 118(3), 157-206.
- Autor, David, Maestas, Nicole, Mullen, Kathleen, and Strand, Alexander (2015). Does delay cause decay? The effect of administrative decision time on the labor force participation and earnings of disability applicants. NBER Working Paper No. 20840.
- Baicker, Katherine, et al. (2013). The Oregon experiment – effects of Medicaid on clinical outcomes. *New England Journal of Medicine*, 368(18), 1713-1722.
- Bound, John (1989). The health and earnings of rejected disability insurance applicants. *The American Economic Review*, 79(3), 482-503.
- Card, David, Dobkin, Carlos, and Maestas, Nicole (2009). Does Medicare save lives? *Quarterly Journal of Economics*, 124(2), 597-636.
- Chen, Susan, and van der Klaauw, Wilbert (2008). The work disincentive effects of the disability insurance program in the 1990s. *Journal of Econometrics*, 142(2), 757-784.
- Coe, Norma, Lindner, Stephan, Wong, Kendrew, and Wu, April Yanyuan (2013). How do the disabled cope while waiting for SSDI? Center for Retirement Research at Boston College Working Paper 2013-12.
- Currie, Janet and Gruber, Jonathan (1996). Saving babies: The efficacy and cost of recent changes in the Medicaid eligibility of pregnant women. *Journal of Political Economy*, 104(6), 1693-1296.
- Finkelstein, Amy et al. (2012). The Oregon health insurance experiment: Evidence from the first year. *Quarterly Journal of Economics*, 127(3), 1057-1106.
- French, Eric and Song, Jae (2014). The effect of disability insurance receipt on labor supply. *American Journal of Economics: Economic Policy*, 6(2), 291-337 .
- Garcia-Gomez, Pilar and Gielen, Anne (2014). Health effects of containing moral hazard: evidence from disability insurance reform. IZA Working Paper 8386.
- Gelber, Alexander, Moore, Timothy and Strand, Alexander (2018). DI benefits save lives. Goldman School of Public Policy Working Paper.
- Gelber, Alexander, Moore, Timothy and Strand, Alexander (2018). The Effect of Disability Insurance Payments on Beneficiaries' Earnings. *American Economic Journal: Economic Policy*, 9(3), 229-261.
- Gross, Tal and Notowidigdo, Matthew (2011). Health insurance expansion and the consumer bankruptcy decision: Evidence from expansions of Medicaid. *Journal of Public Economics*, 95(7-8), 767-778.

- Livermore, G., Stapleton, D., and Claypool, H. (2009). Health Insurance and Health Care Access Before and After SSDI Entry. New York: The Commonwealth Fund.
- Maestas, Nicole, Mullen, Kathleen, and Strand, Alexander (2013). Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. *American Economic Review*, 103(5), 1797-1829.
- Michalopoulos, Charles, et al. (2011). The Accelerated Benefits Demonstration and Evaluation Project: Impacts on Health and Employment at Twelve Months. New York: MDRC.
- Moore, Timothy (2015). The employment effects of terminating disability benefits. *Journal of Public Economics*, 124, 30-43.
- National Center for Health Statistics. Survey Description, National Health Interview Survey, 2005. Hyattsville, Maryland. 2006.
- Social Security Administration (2016). Monthly Statistical Snapshot, February 2018. Accessed 4/12/2018 at https://www.ssa.gov/policy/docs/quickfacts/stat_snapshot/
- Social Security Administration, Office of the Inspector General (2009). Congressional Response Report: Impact of the Social Security Administration's Claims Process on Disability Beneficiaries. A-01-09-29084. Available at <https://oig.ssa.gov/sites/default/files/audit/full/pdf/A-01-09-29084.pdf>
- The Social Security Act, Title II, § 223(d)(1)(A), 42 U.S.C. § 423 (d)(1)(A) and Title 16, § 1614(a)(3)(A), 42 U.S.C. § 1382c(a)(3)(A).
- Wixon, B. and Strand, A., 2013. Identifying SSA's sequential disability determination steps using administrative data. Research and Statistics Note No. 2013-01. Available at <http://www.ssa.gov/policy/docs/rsnotes/rsn2013-01.html>.

APPENDIX

Figure A1. Distribution of Expected Wait Time Instrument



Notes: Figure displays the fraction of applicants whose expected wait time instrument falls into 10-day initial wait bins, with the top bin containing those whose expected waits were 150 days or longer.

Table A.1 Effect of Accountability Incentives on Special Education Placement by Percentage in Special Education

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Reading Level 2*Incentive	0.0113** (0.00420)		0.00530 (0.00387)		0.0179** (0.00412)		0.0142** (0.00389)	
Math Level 2*Incentive	-0.0156** (0.00486)		-0.0163** (0.00415)		-0.00755 (0.00471)		-0.0104** (0.00393)	
Reading Level 3*Incentive	0.0228** (0.00378)		0.0223** (0.00364)		0.0312** (0.00423)		0.0290** (0.00392)	
Math Level 3*Incentive	-0.00232 (0.00272)		-0.00213 (0.00314)		-0.00117 (0.00242)		0.000735 (0.00283)	
Reading Level 2*Distance*Incentive		0.0111* (0.00443)		0.00567 (0.00409)		0.0187** (0.00437)		0.0159** (0.00414)
Math Level 2*Distance*Incentive		-0.0186** (0.00533)		-0.0179** (0.00459)		-0.00914 (0.00522)		-0.0103* (0.00443)
Reading Level 3*Distance*Incentive		0.0244** (0.00427)		0.0261** (0.00416)		0.0354** (0.00475)		0.0349** (0.00450)
Math Level 3*Distance*Incentive		-0.00579 (0.00327)		-0.00237 (0.00387)		-0.00392 (0.00290)		0.00182 (0.00361)
Reading Prior Pass*SWD Incentive	0.00882 (0.00573)	0.00860 (0.00575)			0.0132* (0.00573)	0.0128* (0.00574)		
Math Prior Pass*SWD Incentive	0.0157 (0.00955)	0.0174 (0.00982)			0.0230* (0.00970)	0.0249* (0.01000)		
Reading Prior Score*SWD Incentive			0.00329 (0.00308)	0.00361 (0.00312)			0.00551 (0.00312)	0.00585 (0.00316)
Math Prior Score*SWD Incentive			0.00947** (0.00315)	0.0100** (0.00329)			0.0113** (0.00323)	0.0119** (0.00338)
N	625785	625785	625785	625785	573952	573952	573952	573952
Sample	>= 12.5	>= 12.5	>= 12.5	>= 12.5	<12.5	<12.5	<12.5	<12.5

Notes: This table displays results from 8 linear probability models following Equation (4), one in each column, in which being in special education with a malleable diagnosis is the dependent variable. Each model includes demographic controls, year-by-grade fixed effects, school-level fixed effects, and the main effects of school incentives and prior student performance. The sample is divided by whether the school had more or less than 12.5 percent of its student body in special education. Standard errors in parentheses are clustered at the school level. * denotes significance at the 0.05 level, ** at the 0.01 level.

Table A.2 Effect of Accountability Incentives on Malleable Diagnoses, Including Those with Non-Malleable Diagnoses in Sample

	(1)	(2)	(3)	(4)
Reading Level 2*Incentive	0.0115** (0.00274)		0.00685** (0.00259)	
Math Level 2*Incentive	-0.0127** (0.00320)		-0.0139** (0.00279)	
Reading Level 3*Incentive	0.0220** (0.00266)		0.0204** (0.00253)	
Math Level 3*Incentive	-0.00289 (0.00183)		-0.00210 (0.00203)	
Reading Level 2*Distance*Incentive		0.0115** (0.00290)		0.00744** (0.00274)
Math Level 2*Distance*Incentive		-0.0146** (0.00351)		-0.0150** (0.00308)
Reading Level 3*Distance*Incentive		0.0241** (0.00300)		0.0241** (0.00290)
Math Level 3*Distance*Incentive		-0.00570** (0.00220)		-0.00247 (0.00250)
Reading Prior Pass*SWD Incentive	0.0193** (0.00438)	0.0194** (0.00442)		
Math Prior Pass*SWD Incentive	0.0123 (0.00638)	0.0140* (0.00661)		
Reading Prior Score*SWD Incentive			0.00964** (0.00235)	0.0101** (0.00238)
Math Prior Score*SWD Incentive			0.00652** (0.00233)	0.00661** (0.00240)
N	1237846	1237846	1237846	1237846

Notes: This table displays results from 4 linear probability models following Equation (4), one in each column, in which being in special education with a malleable diagnosis is the dependent variable. Each model includes demographic controls, year-by-grade fixed effects, school-level fixed effects, and the main effects of school incentives and prior student performance. Standard errors in parentheses are clustered at the school level. * denotes significance at the 0.05 level, ** at the 0.01 level.

Table A.3 Descriptive Statistics for Students with Non-Malleable Diagnoses

	Mean	SD
Special Education	1	0
Prior Pass Reading	0.351838	0.47755
Prior Pass Math	0.399542	0.489811
Native American	0.03758	0.19018
Asian	0.008678	0.092753
Hispanic	0.075343	0.263948
Black	0.445011	0.496974
White	0.406301	0.491149
Other	0.027087	0.162338
Female	0.345264	0.475461
Free or Reduced-Price Lunch	0.714748	0.451541
School failed in math	0.281229	0.449605
School failed in reading	0.231421	0.421746
Math Score	-0.83819	1.111153
Prior math score	-0.87208	1.185548
Reading Score	-0.70621	1.198559
Prior reading score	-0.79773	1.21959
N	38,026	

Notes: Table presents descriptive statistics for the sample of students with non-malleable diagnoses. These students are excluded from the main analysis sample. Current and prior test scores are presented in standard deviation units.

Table A.4 Effect of Accountability Incentives on Special Education Placement – Non-Title I Schools

	(1)	(2)	(3)	(4)
Reading Level 2*Incentive	-0.0492** (0.0156)		-0.0418** (0.0130)	
Math Level 2*Incentive	-0.0149 (0.00765)		-0.0207** (0.00655)	
Reading Level 3*Incentive	-0.00528 (0.00643)		-0.0156 (0.00866)	
Math Level 3*Incentive	0.00594 (0.00531)		0.00508 (0.00463)	
Reading Level 2*Distance*Incentive		-0.0566** (0.0166)		-0.0490** (0.0138)
Math Level 2*Distance*Incentive		-0.0182* (0.00893)		-0.0232** (0.00765)
Reading Level 3*Distance*Incentive		-0.00867 (0.00730)		-0.0205* (0.0100)
Math Level 3*Distance*Incentive		0.00338 (0.00598)		0.00585 (0.00552)
Reading Prior Pass*SWD Incentive	-0.0556* (0.0245)	-0.0547* (0.0246)		
Math Prior Pass*SWD Incentive	0.0289 (0.0198)	0.0323 (0.0202)		
Reading Prior Score*SWD Incentive			-0.0172 (0.00884)	-0.0177* (0.00893)
Math Prior Score*SWD Incentive			0.0156* (0.00767)	0.0164* (0.00790)
N	1047084	1047084	1047084	1047084

Notes: This table displays results from 4 linear probability models following Equation (4), one in each column, in which being in special education with a malleable diagnosis is the dependent variable. Each model includes demographic controls, year-by-grade fixed effects, school-level fixed effects, and the main effects of school incentives and prior student performance. Standard errors in parentheses are clustered at the school level. * denotes significance at the 0.05 level, ** at the 0.01 level.

Table A.5 Effects of Special Education on Gains in Math Scores, Hanushek, Kain, & Rivkin model

	No Student FE	Student FE	N
	(1)	(2)	
All	0.0261** (0.0028)	0.0663** (0.0068)	2,311,728
Diagnosis			
Learning Disabled	0.0635** (0.0052)	0.0930** (0.0107)	157,442
Speech/Language	0.0272** (0.0051)	0.0280* (0.0126)	70,901
Emotional/Behavioral	0.0923** (0.0213)	0.1030 (0.0527)	14,560
Other Health (ADHD)	0.0818** (0.0087)	0.1065** (0.0192)	72,986
Autism	0.0429 (0.0323)	0.0865 (0.0801)	17,071
Previous Achievement Level			
1	0.0563** (0.0061)	0.1991** (0.0223)	68,653
2	-0.0027 (0.0035)	0.1461** (0.0129)	248,366
3	-0.0717** (0.0028)	0.0723** (0.0094)	673,307
4	-0.1672** (0.0081)	0.0297 (0.0163)	320,893

Notes: This table presents the results of 20 models following Equation (6), in which the gain in math z-score is the dependent variable. Standard errors in parentheses are clustered at the school level. A * denotes significance at the 0.05 level and **denotes significance at the 0.01 level.

Table A.6 Effect of Accountability Incentives on Special Education Placement – Alternate Malleable Definition

	(1)	(2)	(3)	(4)
Reading Level 2*Incentive	0.0167** (0.00337)		0.0108** (0.00312)	
Math Level 2*Incentive	-0.0118** (0.00397)		-0.0144** (0.00332)	
Reading Level 3*Incentive	0.0281** (0.00328)		0.0264** (0.00313)	
Math Level 3*Incentive	0.000585 (0.00220)		0.000593 (0.00247)	
Reading Level 2*Distance*Incentive		0.0165** (0.00357)		0.0115** (0.00331)
Math Level 2*Distance*Incentive		-0.0147** (0.00438)		-0.0151** (0.00372)
Reading Level 3*Distance*Incentive		0.0306** (0.00372)		0.0312** (0.00360)
Math Level 3*Distance*Incentive		-0.00250 (0.00264)		0.00151 (0.00310)
Reading Prior Pass*SWD Incentive	0.0261** (0.00543)	0.0262** (0.00555)		
Math Prior Pass*SWD Incentive	0.0163 (0.00854)	0.0177* (0.00888)		
Reading Prior Score*SWD Incentive			0.0127** (0.00307)	0.0133** (0.00311)
Math Prior Score*SWD Incentive			0.00841** (0.00308)	0.00873** (0.00321)
N	1199737	1199737	1199737	1199737

Notes: This table displays results from 4 linear probability models following Equation (4), one in each column, in which being in special education with a malleable or autism diagnosis is the dependent variable. Each model includes demographic controls, year-by-grade fixed effects, school-level fixed effects, and the main effects of school incentives and prior student performance. Standard errors in parentheses are clustered at the school level. * denotes significance at the 0.05 level, ** at the 0.01 level.

Table A7. Effects of Wait Time on Program Status, with Survey Weights

	Benefit Status at Survey					
	Ever SSDI (1)	Current (2)	Suspended (3)	Terminated (4)	Never Benefits (5)	Any Re- consideration (6)
Panel 1 – No Lag						
Wait	-0.000020 (0.00092)	0.0021+ (0.0012)	0.00021 (0.00027)	-0.0019** (0.00066)	-0.00042 (0.0011)	-0.0018 (0.0013)
F-statistic	24.04	24.04	24.04	24.04	24.04	24.04
N	2131	2131	2131	2131	2131	2131
Panel 2 – 1 Month Lag						
Wait	0.00088 (0.0015)	0.0039** (0.0015)	-0.00017 (0.00041)	-0.0014 (0.00089)	-0.0024+ (0.0014)	-0.0028+ (0.0016)
F-statistic	23.22	23.22	23.22	23.22	23.22	23.22
N	2064	2064	2064	2064	2064	2064

Notes: Table reports the estimates of the effect of wait time on outcomes, from 12 IV regressions. The second panel controls for 1-month lagged wait time. Standard errors are clustered at the state level. Kleibergen-Paap F-statistic on the instrument is reported at the bottom of each panel. + denotes significance at the 0.1 level, * at the 0.05 level, and ** at the 0.01 level.

Table A8. Effects of Wait Time on Health and Well-Being, with Survey Weights

	Activity Limitations (1)	Functional Limitations (2)	Good or Better Health (3)	Bed Days (4)	BMI (5)	Mental Health Symptoms (6)	Health Care Access Index (7)	SNAP (8)	<= 100% FPL (9)	Working Last Week (10)
Panel 1 – No Lag										
Wait	0.010* (0.0041)	0.0070 (0.0051)	-0.00068 (0.0015)	-0.081 (1.06)	0.042+ (0.025)	0.0018 (0.0018)	-0.00060 (0.0012)	-0.00029 (0.0011)	0.00045 (0.0015)	-0.00075 (0.0010)
F-statistic	24.04	24.04	23.41	7.32	6.92	7.25	5.74	22.66	23.08	22.70
N	2131	2131	2127	1131	969	1122	1079	2119	1717	2120
Panel 2 – 1 Month Lag										
Wait	0.017** (0.0055)	0.0061 (0.0068)	-0.0019 (0.0018)	-0.14 (1.91)	0.051 (0.039)	0.0037 (0.0034)	-0.00034 (0.0020)	0.00058 (0.0017)	0.0025 (0.0016)	-0.00050 (0.0012)
F-statistic	23.22	23.22	22.43	4.60	4.72	4.56	3.37	21.50	21.10	20.99
N	2064	2064	2060	1089	933	1080	1038	2052	1662	2054

Notes: Table reports the estimates of the effect of wait time on outcomes, from 20 IV regressions. The second panel controls for 1-month lagged wait time. Standard errors are clustered at the state level. Kleibergen-Paap F-statistic on the instrument is reported at the bottom of each panel. + denotes significance at the 0.1 level, * at the 0.05 level, and ** at the 0.01 level.

Table A9. Effect of Wait Time on Components of Health Care Access Index

	Forewent due to cost:								
	Dentist Past 12mo (1)	Prescription Medicine (2)	Mental health Care (3)	Dental Care (4)	Medical Care (5)	Delayed for Cost (6)	Usual Place for Care (7)	No Insurance (8)	> \$500 Out of Pocket (9)
Panel 1 – No Lag									
Wait	0.00035 (0.0014)	-0.00064 (0.0015)	-0.0025 (0.0017)	0.0011 (0.0015)	-0.00074 (0.00098)	-0.00095 (0.00088)	0.00020 (0.0012)	-0.0013 (0.0012)	0.00029 (0.00091)
F- statistic	8.60	8.07	8.00	8.07	33.30	33.25	8.476	31.61	30.04
N	1116	1121	1119	1120	2128	2129	1108	2121	2087
Panel 2 – 1 Month Lag									
Wait	-0.0035 (0.0030)	0.0011 (0.0023)	-0.0029 (0.0032)	0.0017 (0.0024)	-0.0013 (0.0016)	-0.0012 (0.0014)	0.0011 (0.0017)	-0.0011 (0.0019)	0.0018 (0.0015)
F- statistic	7.52	7.44	7.33	7.53	41.55	41.17	7.781	39.16	37.17
N	1075	1079	1077	1078	2061	2062	1066	2054	2020

Notes: Table reports the estimates of the effect of wait time on outcomes, from 18 IV regressions. The second panel controls for 1-month lagged wait time. Standard errors are clustered at the state level. Kleibergen-Paap F-statistic on the instrument is reported at the bottom of each panel. + denotes significance at the 0.1 level, * at the 0.05 level, and ** at the 0.01 level.



An evaluation of the Mellon Mays Undergraduate Fellowship's effect on PhD production at non-UNCF institutions[☆]



Sarah J. Prenovitz, Gary R. Cohen^{*}, Ronald G. Ehrenberg, George H. Jakubson

Cornell University, Department of Economics, 465 Uris Hall, Ithaca, NY 14853-7601, United States

ARTICLE INFO

Article history:

Received 17 November 2014
Revised 13 April 2016
Accepted 14 April 2016
Available online 4 May 2016

JEL classification:
I23

Keywords:

Mellon Mays Undergraduate Fellowship
Program evaluation
Education
PhD production

ABSTRACT

The Mellon Mays Undergraduate Fellowship Program (MMUF) encourages underrepresented minority (URM) students to pursue PhD study with an eye toward entering academia. Fellows have completed PhDs at high rates relative to other students, but they are selected for their interest and potential. In this paper we use restricted access data from the Mellon Foundation and the National Science Foundation's *Survey of Earned Doctorates* to investigate the effect of the MMUF on PhD completions by URM students who graduate from participating institutions. We find no evidence that participation in the program causes a statistically significant increase in the numbers of PhDs completed by URM students, and increases greater than about one PhD per institution per cohort lie outside a 95% confidence interval of our estimates. This suggests that at least some of the PhDs completed by participants would have occurred without the program.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Colleges and universities seek to diversify their faculty along several dimensions, but find few underrepresented minorities¹ in their hiring pool, with the problem worse in some fields than others. This is a manifestation of what is often referred to as the pipeline problem. Relatively few minorities pursue graduate study in many disciplines in the humanities, social sciences, physical sciences, or life sciences. If individuals do not enter PhD programs in a given field they will not emerge from the other end of the pipeline as potential faculty members.

[☆] Cornell Higher Education Research Institute (CHERI). CHERI receives financial support from the Andrew W. Mellon Foundation but the conclusions we express here are strictly our own. The use of NSF data does not imply NSF endorsement of the research, research methods, or conclusions contained in this paper.

^{*} Corresponding author. Tel.: +1 440 669 2594.

E-mail address: garyrcohen@gmail.com, grc64@cornell.edu (G.R. Cohen).

¹ Underrepresented minorities are defined as those who identify neither as non-Hispanic White nor as Asian.

The Mellon Minority Undergraduate Fellowship Program, since renamed the Mellon Mays Undergraduate Fellowship Program (MMUF), was established in 1988 with the goal of addressing this issue by encouraging underrepresented minorities to pursue graduate study in particular fields, with an eye toward ultimately entering academia. Participating schools select fellows from among their students, coordinate mentoring, and hold regular seminars which emphasize research and graduate school. Fellows receive stipends to allow them to conduct research as undergraduates. They are eligible to attend regional and national conferences at which they can present their own research, learn about that done by other fellows, and network. Fellows can also receive up to \$10,000 in loan repayments.

As of 2014, over 4000 students have participated in the program; 506 have earned PhDs and another 665 PhDs are in progress (Bengochea, 2013). As the program has expanded over time and as most PhD programs take at least 5 years to complete, if not substantially longer, this suggests an extremely high rate of PhD completion by MMUF scholars. A back of the envelope estimate suggests

that about a quarter of MMUF students will eventually complete a PhD, compared with around 4% of under-represented minority students graduating from MMUF institutions in years their school was not participating in the program.²

Anecdotal evidence suggests that the MMUF may play a large role in the ultimate PhD completion of its participants. MMUF administrators report struggling to recruit undergraduate candidates because few students have considered the possibility of entering academia, and fellows cite their research experiences, relationships with mentors, and connections with other fellows as crucial to their decision to pursue a PhD and their ultimate success in completing one (Rose, 2012). However, fellows presumably apply to the program because of their own interest and are selected based on their potential as scholars, so the high rate of PhD completion reflects both this selection and the effects of the program. Indeed, in a 2007 survey, 67% of current and former fellows responded that they would have or might have aspired to earn a PhD absent the program (Rose, 2012). The MMUF may still help students turn these goals into reality, and inspire those who would not have otherwise considered an academic career to explore one, but fellows are probably quite different from other students in their underlying propensity to complete a PhD. It also may be the case that students who have already completed the program overstate their chances of pursuing PhDs in its absence. Without a doubt some current and former fellows would have aspired to earn PhDs in the program's absence, and so it is useful to determine to what extent fellows' high PhD completion rate is a result of the program.

We address this issue by estimating the effect of a school's MMUF participation per se and the intensity of participation on the number and rate of PhD completions by under-represented minority (URM) students.³ By

investigating the outcomes of all URM students at an institution we are able to avoid this sample selection problem, and address the effect of the program on its medium-run goal. Institutions have joined the program gradually over time, and this allows us to control for time trends and cross-institution variation using institution and year fixed effects. However, due to the lengthy nature of PhD programs we do not observe the completions of many of those who will eventually earn a PhD, especially in later cohorts. Time to degree is in general longer for URM students than for other students, exacerbating the problem. We estimate the size of this truncation using data from those who graduated in the early years of our dataset, and conduct the bulk of our analyses using this adjusted data.

While our focus is on a single fellowship, many other programs share the broad goals and methods of the MMUF. To our knowledge these programs have not been evaluated in the economics literature, but several prior studies in the education literature have explored their effects. Most of this work has been purely correlational, which is problematic as students who participate are quite different from those who do not. Other analyses have used propensity score matching to construct an appropriate control group of students, but participants may still differ from non-participants in important but unobservable ways (e.g. Eagan et al., 2013). We contribute to this literature by using a design that allows us to avoid the issue of student selection, addressing problems of truncation in degree data, and analyzing a program whose causal effects are unknown.

We estimate the average effect of an institution's participation in the MMUF program and find no statistically significant effect of the program when considering only the MMUF schools. These findings persist when we account for truncation and when we add control groups constructed through propensity score matching. We also find no effect of adding an additional fellow or increasing the percentage of URM students who are fellows. This is particularly notable as these estimates may suffer from positive selection bias: institutions are able to move funds from year to year, awarding more fellowships in years with relatively strong applicant pools and fewer in other years. In addition, our confidence intervals rule out a causal effect of more than one additional PhD per cohort on average, with an average cohort size of 4.8 students. Whether that effect is meaningful is open to interpretation. If we assume that about a quarter of fellows will eventually go on to complete PhDs, it suggests the MMUF is supporting some students who would complete PhDs anyway. Because we are evaluating a small program using aggregate data, it is possible the program has an effect that is too small for us to distinguish. It may also be that the MMUF is important to participants in other ways that do not significantly increase the number of PhDs, or that our truncation adjustments do not adequately capture changes in the time to degree over time.

The rest of this paper proceeds as follows. Section 2 provides background and further detail on the program structure and history. Section 3 describes our data and methods. Section 4 presents results and Section 5 discusses these results and our conclusions.

² If we assume that the program selected the same number of fellows in each year for a total of 4000 as of 2014 and their distribution of completion times was the same as non-white non-Asian students who completed a bachelors' degree in a MMUF field in 1985–1989 and went on to complete a PhD, we would expect to observe about half of those PhDs which will be completed within 20 years of graduation. As approximately one in eight MMUF fellows has completed a PhD, this suggests that about a quarter of MMUF fellows will do so eventually. The program has grown over time, which would make this back of the envelope calculation an underestimate. However, fellows also have incentives and supports to complete degrees more quickly, so degrees completed so far may represent a larger proportion of those that will eventually be completed.

³ The Mellon Foundation considers throughput—the entry of its fellows into PhD programs—to be a key metric for assessing the undergraduate components of the program (Bengochea, 2013). While the PhD completion rate of the URM student body as a whole is not an explicit program goal, the best available data for causal analysis limit our focus to completion and to an estimate of the treatment effect on an average URM student at a participant school rather than a Mellon Mays fellow specifically. We believe that this average treatment effect is an appropriate metric for evaluation, as it is closely linked to the pipeline problem that is the *raison d'être* for the MMUF, and any increase in fellows' PhD completion that results from the program should also increase the overall number of URM students completing PhDs.

2. Background and program structure

2.1. Background

Colleges and universities pursue faculty diversity for several reasons. First, if minority faculty members are better at connecting with minority students, either in the classroom or as mentors and role models, their presence might be important to the persistence and graduation rates of minority students. There is some evidence that minority students are more likely to persist in STEM majors if they have an introductory STEM course that is taught by a minority professor (Price, 2010), and that gaps between minority and non-minority community college students in pass rates, grades, and courses dropped are smaller when classes are taught by professors who are minorities themselves (Fairlie, Hoffmann, & Oreopoulos, 2011). Second, to the extent that raw teaching and research potential are distributed throughout the population, hiring underrepresented minorities at a very low rate implies that institutions are losing out on important groups of potential faculty. Finally, diversity may be pursued for its own merits. It can stimulate a dynamic academic atmosphere, enriching the work and lives of all faculty and students; address societal inequalities; or bring academic attention to a wider range of issues that would otherwise be the case.

Institutions that seek to diversify faculty are constrained by the small number of underrepresented minorities completing PhDs. In 2011, 6.14% of the US citizens or permanent residents earning a PhD reported that they were Black, while 6.3% reported that they were Hispanic (National Science Foundation, 2012). These numbers are substantially higher than at the inception of the MMUF (4.8% and 3.6% in 1985 respectively), but still quite low relative to the US population, which was 12.3% Black and 16.7% Hispanic in 2011 (United States Census Bureau, 2011). There is also substantial variation across fields, with Black students earning 13% of PhDs in education but only 3% of those in physical sciences, and Hispanic students earning 8% of PhDs in social sciences and 4.5% of those in the physical sciences (NSF, 2012). While departments in some fields might find a diverse range of job candidates others are still constrained in their ability to hire from underrepresented minority groups.

2.2. The Mellon Minority/Mays Undergraduate Fellowship Program

The MMUF program began with eight institutions, which joined the program in late 1988 and recruited their first fellows in the spring of 1989. Additional cohorts joined in 1989, 1992, 1996, 2000, and 2007. A group of Historically Black Colleges and Universities has participated since 1989 through a consortium administered by the United Negro College Fund (UNCF).⁴ Not counting this

consortium, 42 institutions participated in the program in 2014 (Mellon Mays Undergraduate Fellowship, 2013). A table of the institutions in our sample and the year they joined the program appears in Appendix Table A1.

Participating institutions select fellows, generally targeting students in the spring of their sophomore year. Schools are provided with funding for up to five fellows per year, though they are able to select more fellows in some years by moving funds from one school year to another, or if students who were previously selected drop out of the program. In the early years of the program fellowships were restricted to those belonging to underrepresented minority groups. However, in response to concerns from participating institutions about the legality and ethics of affirmative action and other race-based programs, eligibility was extended in 2003 to students of all backgrounds who were committed to the program's goal of increasing the presence of underrepresented minorities in academia (Mellon Foundation, 2003). In addition to supporting the diversity goals of the program, fellows must be pursuing a major in one of the Mellon-designated fields. These span the humanities, social sciences, and natural sciences, but do not include all majors. A list of the fields for 2000 and 2008 is included in Appendix Table A2.

Students apply directly to the fellowship program at their institution. Although each institution has considerable discretion in evaluating applicants, they are asked to consider the student's field and either minority status or commitment to the program's goals, as well as academic promise, interest in an academic career, and potential as a mentor. Once selected as fellows, students work with mentors and attend seminars at their home institution. Because of the decentralized nature of the program, each participating school decides how to implement the mentorship and seminar components. Mentors are intended to act as graduate school advisers—much as a pre-law or pre-med adviser would—and to oversee the student's independent research. Seminars are in general focused on research and preparation for graduate school, and are intended also to allow students to form a group identity. Fellows also receive stipends both during the school year and over the summer to allow them to focus on research rather than paid work, and to potentially allow for fieldwork or study at another institution over the summer. The MMUF administers regional and national conferences at which students can present work, be exposed to the work done by other fellows, and network with current and former fellows.

After college graduation, fellows who attend graduate school in a designated field can be eligible to participate in seminars and conferences, apply for grants expressly for former undergraduate fellows, and receive loan forgiveness. The seminars and conferences include an annual conference similar to that attended by undergraduates as well as programs focused on writing grants and dissertations. There are also retreats for those in the dissertation-writing phase. Loans taken out for undergraduate and graduate study are eligible for loan forgiveness, up to

⁴ The UNCF is a consortium of 37 private Historically Black Colleges and Universities (HBCUs) and among the most well-known of these institutions are Clark Atlanta, Fisk, Morehouse and Spelman. The consortium is permitted to choose up to 25 fellows a year from across their member institutions. As of the Spring of 2015, 614 fellows had been selected. Ninety

of these fellows had completed PhDs and 71 more were enrolled in PhD programs at that time. Our analyses exclude the UNCF consortium for reasons described in the methods section.

a total of \$10,000 if the student completes a PhD in a Mellon-designated field. Loan forgiveness is only available to those who attend a PhD or terminal Masters' program in one of the designated fields, and requires the fellow to begin his or her program within about 3 years of graduation or submit an appeal.

3. Data and methods

3.1. Data and sample

Our analyses use data from the *Integrated Post-Secondary Education Data System* (IPEDS), a restricted access version of the National Science Foundation's *Survey of Earned Doctorates* (SED), and restricted access administrative data from the Mellon Foundation. The IPEDS includes institution-level information on enrollment, costs and finances, faculty, and other characteristics for all colleges and universities in the United States that receive federal funding. Data is provided by institutions, and has been collected in 1987 and then annually since 1989. The *Higher Education General Information System* (HEGIS), the predecessor to IPEDS, includes data for earlier years, dating back to 1966. Many of the variables in the HEGIS data are the same or similar to those in IPEDS, but HEGIS includes less information and was collected less frequently. From these systems we obtain institution-level information on the number of students completing bachelors' degrees by race/ethnicity, gender, and field from 1985 through 2005. We also use a larger set of institutional characteristics, drawn from the IPEDS Delta Cost Project data, in order to construct matched comparison groups.

The *Survey of Earned Doctorates* (SED) is an annual census of PhD completers in the United States, sponsored by six federal entities and administered since 1957. The SED achieves a high response rate, with 92% of those earning PhDs responding in 2012, although item response rates are somewhat lower (National Science Foundation, 2013). It includes information on demographics, undergraduate and graduate study, and career plans. We use data from the SED for those who completed a PhD between 1985 and 2011. From this population we count the number of individuals who have completed a PhD by undergraduate institution, year of bachelors' degree, minority status, field, and gender. We also obtain counts of the number completing a PhD a given number of years after the bachelors' degree by subgroup and year of bachelors' degree. This data on the distribution of times to PhD completion for the early cohorts in our sample is used to adjust for the fact that later cohorts have fewer years to complete a PhD, and thus their numbers of PhD completers are understated due to sample truncation.

The Mellon Foundation's data provides counts of MMUF program fellows at each participant institution in each year. The distribution of fellows by year and school is depicted in Fig. 1. Although most institutions had 3–6 fellows in most years of participation there is considerable variation in the number of fellows. In general, the smallest schools in terms of URM enrollment are more likely to have fewer than 5 fellows in multiple years. We use this data to calculate the 'dosage' of the program within the

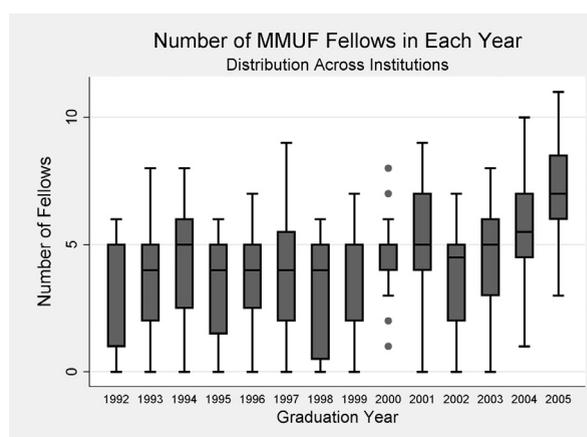


Fig. 1. Displays the distribution of the number of fellows across institutions. For each year the smallest and largest cohorts are marked, as well as the 25th, 50th, and 75th percentiles. In 1992 the smallest cohort was 0 fellows, the 25th percentile 1 fellow, the 75th 5 fellows, and the largest 6 fellows. In 1993 the smallest cohort had 0 fellows, the 25th percentile 2 fellows, the 50th 4, the 75th 5, and the largest 8 fellows. In 1994 the smallest cohort had 0 fellows, the 25th percentile 3, the 50th 5, the 75th 6, and the largest 8 fellows. In 1995 the smallest cohort had 0 fellows, the 25th percentile 2, the 50th 4, the 75th 6, and the largest 7 fellows. In 1996 the smallest cohort had 0 fellows, the 25th percentile 3, the 50th 4, the 75th 5, and the largest 7 fellows. In 1997 the smallest cohort had 0 fellows, the 25th percentile 3, the 50th 4, the 75th 6, and the largest 9 fellows. In 1998 the smallest cohort had 0 fellows, the 25th percentile 1, the 50th 4, the 75th 5, and the largest 6 fellows. In 1999 the smallest cohort had 0 fellows, the 25th percentile 2, the 75th 5, and the largest 7 fellows. In 2000 there were two outliers with small values (1 and 2) and two with large values (7 and 8). Among the rest of the sample, the smallest cohort had 3 fellows, the 25th percentile 3, the 75th 5, and the largest 6 fellows. In 2001 the smallest cohort had 0 fellows, the 25th percentile 4, the 50th 5, the 75th 7, and the largest 9 fellows. In 2002 the smallest cohort had 0 fellows, the 25th percentile 2, the 50th 4, the 75th 5, and the largest 7 fellows. In 2003 the smallest cohort had 0 fellows, the 25th percentile 3, the 50th 5, the 75th 6, and the largest 8 fellows. In 2004 the smallest cohort had 1 fellow, the 25th percentile 3, the 50th 6, the 75th 7, and the largest 10 fellows. In 2005 the smallest cohort had 3 fellows, the 25th percentile 6, the 50th 5, the 75th 7, and the largest 11 fellows. Source: Administrative data provided by the Mellon Foundation.

overall URM population of each cohort at each institution. We then use this dosage and the raw numbers of participants in extensions to our base model to account for the fact that we would expect the treatment effect of this relatively small program to be more pronounced—and easier to detect—at institutions where a greater proportion of the URM population participated.

Our analyses focus on the 32 non-UNCF institutions that selected their first fellows by 2005. The UNCF institutions are excluded because program participation at each UNCF school in any particular year is far more varied than at the other U.S. institutions with most UNCF institutions having zero participants in any given year. We are concerned that we would be unable to discern the effect of the program at these schools, and that the participation of a given institution in the program in a particular year may be a strong signal about the propensity of its students in that cohort to attend graduate school. We limit our analysis to cohorts that graduated by 2005, as more recent graduates had completed relatively few PhDs by 2011.

Table 1
 Characteristics of the non-UNCF institutions participating in the MMUF program by 2005.

	N	Mean	SD
Public control	32	0.125	0.336
Highest degree			
Bachelors	32	0.219	0.420
Doctorate	32	0.594	0.499
Masters/first professional	32	0.188	0.397
Characteristics as of 1987			
Enrollment	32	8901.4	7768.5
Tuition and fees per student	32	8834.22	3335.80
Percent of students who are undergraduates	32	0.734	0.217
1985 Graduates			
URM BAs	32	214.1	495.9
non-URM BAs	32	831.7	610.6
Proportion of BAs awarded to URM students	32	0.145	0.202
PhD Completion Rate	32	0.106	0.078
2005 Graduates			
URM BAs	31	333.1	269.6
non-URM BAs	31	989.5	702.5
Proportion of Bas awarded to URM students	31	0.229	0.12
PhD completion rate	31	0.017	0.03
Full sample, unadjusted			
URM PhDs—arts and sciences	693	5.94	5.43
URM PhDs—arts and sciences and engineering	693	6.47	5.91
URM PhDs—all fields	693	7.30	6.47
Full sample, simple truncation adjustment			
URM PhDs—arts and sciences	693	7.37	6.58
URM PhDs—arts and sciences and engineering	693	7.99	7.12
URM PhDs—all fields	693	9.20	7.92
Full sample, 10 year truncation adjustment			
URM PhDs—arts and sciences	693	6.94	6.16
URM PhDs—arts and sciences and engineering	693	7.66	6.84
URM PhDs—all fields	693	8.49	7.35

Most institutions in our sample are privately controlled (Table 1). A slight majority grant doctorates, with the rest about evenly split between those that only grant bachelor's degrees and those that grant masters or first professional degrees. In 1985, before the MMUF began, the average institution produced slightly over 1000 bachelors' degrees, about 14% of which went to URM students. About 11% of graduates went on to complete a PhD by 2011, with rates fairly similar for URM and non-URM students. By 2005 the average institution produced about 1300 bachelor's degrees, with about 23% going to URM students. Only 1.7% of these graduates completed a PhD by 2011, which is unsurprising given that they only had 6 years to do so.

We focus primarily on the sample of URM students, based on the assumption that the MMUF has the potential to affect PhD completions for those students who are eligible to participate, but not those who are not eligible. In some analyses we include PhD completions for non-minority students as a control variable. While the introduction of a new fellowship opportunity could decrease competition for existing programs, change campus culture, or inspire the peers of fellows to pursue a different path, we do not expect these factors to be large, particularly as the program is relatively small, and its benefits are restricted to fellows. To the extent that the benefits of the MMUF spill over to those who are not eligible to participate, our estimates would understate the full effect of the MMUF on PhD completion.

Our initial plan was to restrict the analysis to PhD completions in MMUF fields, based on similar reasoning. We were forced to abandon this plan for several reasons. First, IPEDS reports only one major per BA completer until 2000, and two in later years. Thus students with more than one major before 2000, or more than two after, could be eligible for the program but not be identifiable in the data as being eligible. Second, a sizable number of students switch fields between BA and PhD. Third, although more recent data contains detailed information on undergraduate majors, HEGIS and IPEDS used a very broad coding scheme for completers' fields (two-digit CIP) until 1996. As a result we are unable to distinguish some of the MMUF fields in these early data. Finally, even with perfect data we would not be able to define which students were eligible for the program based on fields, both because institutions had some discretion to decide whether a field closely related to a MMUF field was eligible, and because documentation on which fields were eligible before 2000 does not exist. Instead we restrict the sample by broad categories of fields—arts and sciences; arts, sciences, and engineering; and all fields—rather than using a more specific definition of eligible fields. All MMUF fields from 2000 and 2008 fall into arts and sciences, which includes humanities, social sciences, life sciences, and physical sciences. The arts, sciences, and engineering group adds engineering fields. The all fields category includes all BAs or doctorates, including those in arts, humanities,

and engineering, as well as fields such as education and business.

3.2. Method

Using the Survey of Earned Doctorates, we amass data on the number of graduates of a given institution in a given year who have since gone on to earn a PhD.⁵ This is done separately for minority and non-minority students, by the broad field groups described above. We define $BASj_{it}$ as the number of individuals in group J (URM or non-URM) who completed bachelor's degrees at institution i in year t , and $PhDs_{jt}$ as the number of those individuals who completed a PhD by 2011, when our information on PhD completion ends.

In our baseline specification we regress the PhD completion count for URM students ($PhDsM_{it}$) on the count of BAs awarded to that cohort (BAS_{it}) and whether the institution was participating in the program when that cohort was eligible to be selected to participate (MMF_{it}). We also include graduation year fixed effects (T_t) in order to control for variations over time in the number of bachelors' graduates completing PhDs nationally, and institution fixed effects (I_i). Because the dependent variable is a count, we estimate a negative binomial model,⁶ assuming an exponential functional form so that the mean of $PhDsM_{it}$ is given by:

$$E(PhDsM_{it}) = I_i \exp(\beta_1 MMF_{it} + \beta_2 BAS_{it} + T_t) \quad (1)$$

We also estimate the baseline equation with the addition of the count for non-minority students ($PhDsNM_{it}$) included as an explanatory variable, but we find that this does not significantly alter our findings and so exclude it from later analyses. Standard errors are clustered at the institution level.

4. Results

4.1. Baseline estimates

Results from the baseline specification are displayed in Table 2. Despite the benefits of the MMUF felt by its participants, the model is unable to detect any impact of the program on the PhD production of URM graduates. We find no significant effect of the program on PhD completions, and point estimates are mostly less than zero, suggesting small decreases in the number of PhDs completed. Using the negative binomial model, an increase in PhD completions in the arts and sciences larger than 1.001 PhDs per participating school per cohort lies outside a 95% confidence interval. By comparison, the OLS model allows an increase as large as 1.47 within the 95% confidence interval. Results are similar when we include the non-minority PhD

⁵ We also study the PhD completion rate, defined as the proportion of BAs from a given institution in a given year who go on to earn PhDs during our sample. We find that models using PhD counts are easier to interpret, but rate models are discussed as a robustness check, and full results are available in appendix B.

⁶ Our tests indicate overdispersion, so we favor negative binomial regression. Poisson estimation yields very similar results and so is not reported here. Table 2 includes an OLS model for comparison.

Table 2

Effect of MMUF participation on URM PhD production—model comparison.

	Model	OLS	Negative binomial
(a)	A&S	0.466 (0.514)	−0.151 (0.588)
(b)	A&S + Eng.	0.343 (0.515)	−0.273 (0.567)
(c)	All fields	0.525 (0.508)	−0.141 (0.564)

Notes: Six models are reported: for each model, the dependent variable is the number of PhD completions among those non-white, non-Asian students who graduated from an institution in a particular year, with degrees in a particular group of fields as indicated by (a), (b), and (c). All models include the BA completion count for that cohort, as well as year and institution fixed effects. Coefficients are reported for the OLS model and marginal effects are reported for the negative binomial model. Standard errors in parentheses are clustered by institution.

A * indicates significance at the 5% confidence level, and a † the 1% confidence level.

Table 3

Effect of MMUF participation on URM PhD production—unadjusted model.

		(1)	(2)
(a)	A&S	−0.151 (0.588)	−0.222 (0.561)
(b)	A&S + Eng.	−0.273 (0.567)	−0.389 (0.552)
(c)	All fields	−0.141 (0.564)	−0.221 (0.550)
	White and Asian PhDs		✓

Notes: Marginal effects from six negative binomial models are reported. For each model, the dependent variable is the number of PhD completions among those non-white, non-Asian students who graduated from an institution in a particular year, with degrees in a particular group of fields as indicated by (a), (b), and (c). All models include the BA completion count for that cohort as well as year and institution fixed effects. Specification (2) includes the comparable PhD count for white and Asian students. Standard errors in parentheses are clustered by institution.

A * indicates significance at the 5% confidence level, and a † the 1% confidence level.

completion counts (Table 3 columns 2 and 4) and when considering degrees in all fields.

Because more schools have joined the MMUF program over time, and later cohorts suffer greater truncation, our baseline model likely understates the impact of the program. For example, a student who completed his or her bachelor's degree in 2002 has only 9 years to complete a PhD by 2011, the last year of data available to us on PhD completions. This number is below the median time to degree for some fields, and thus will miss more than half of the potential PhDs. Because PhDs in progress at the time of measurement are treated as though they will never be completed, the PhD production rate would appear to be declining over time if it were constant. Then, because the MMUF program is introduced throughout our sample period, participation effects are confounded with truncation effects. URM students take longer on average to complete PhDs, so controlling for the equivalent non-minority count does not eliminate the problem of truncation. Differences in time to degree are illustrated in Fig. 2,

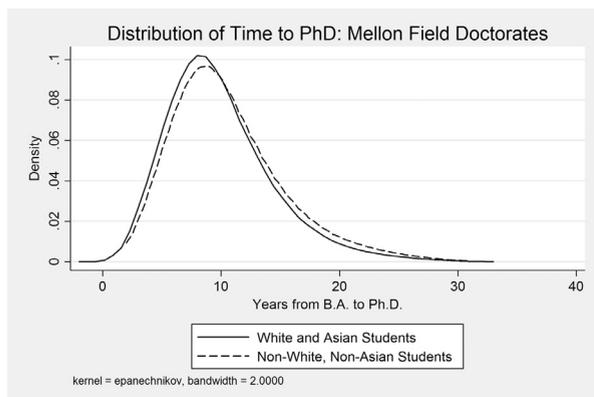


Fig. 2. Displays the distribution of time to PhD separately for underrepresented minorities (Non-White, Nonresults Asian) and White and Asian students. The density for both groups is at first flat near zero then climbs steeply until about 10 years. It then decreases at a slower and decreasing rate. The curve for underrepresented minorities lies to the right of that for White and Asian students.

Notes: Sample includes all US Citizens and Permanent Residents who completed a PhD in one of the MMUF fields between 1980 and 2011, completed a BA in the US, and responded to the Survey of Earned Doctorates. *Source:* Source: Authors' calculations based on the Survey of Earned Doctorates

which presents the distribution of PhD completion times for all SED respondents who completed their BA in the US and are US citizens or permanent residents.⁷ Indeed, Table 3 demonstrates that including information on non-URM PhD completions hardly affects the estimates of program participation at all.

In order to improve this estimate we implement two strategies to adjust for the fact that we do not observe PhDs in progress. The first takes the distribution of time to PhD that prevailed in the first 5 years of our sample and applies it to the remainder of the data. That is, we predict how many of those who have completed bachelor's degrees but are not recorded as having completed PhDs they will eventually complete a PhD, and use that to form our estimate of PhD completions. We do this separately for URM and non-URM students. Students from these early cohorts have at least 20 years post-college to finish their graduate degrees, so truncation is likely to be a much smaller problem. If this method captures truncation patterns accurately it puts all cohorts on an equal footing. The disadvantage of this approach is that it makes the strong assumption that the time-to-PhD distribution is fixed over time.⁸

To address this latter concern, we estimate a second model where we allow the truncation pattern to change over time. We do this by estimating a quadratic model on early cohorts for each number of years from BA y , where t

⁷ This figure understates the difference in degree completion time somewhat, as URM students make up a larger proportion of PhD completions in later cohorts, and members of later cohorts with particularly long times to degree do not appear in the data.

⁸ In fact, formal statistical tests that we conducted suggest that this assumption is not strictly true.

Table 4

Effect of MMUF participation on URM PhD production—truncation adjustments.

		(1)	(2)
(a)	A&S	0.134 (0.737)	0.391 (0.765)
(b)	A&S + Eng.	0.037 (0.731)	0.393 (0.795)
(c)	All fields	0.211 (0.779)	0.518 (0.787)
Simple adjustment		✓	
10-year adjustment			✓

Notes: Marginal effects from six negative binomial models are reported. For each model, the dependent variable is the predicted number of PhD completions among those non-white, non-Asian students who graduated from an institution in a particular year, with degrees in a particular group of fields as indicated by (a), (b), and (c). All models include the BA completion count for that cohort as well as year and institution fixed effects. The simple adjustment is a truncation adjustment under the assumption that the time to PhD pattern from 1985–1989 persists throughout the sample. The 10-year adjustment is a truncation adjustment with a quadratic model in time fit to the first 10 years of data. Standard errors in parentheses are clustered by institution. A * indicates significance at the 5% confidence level, and a † the 1% confidence level.

is the number of years from 1985:

$$Pr(\text{Completeness in } y \text{ years}) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 \quad (2)$$

We then apply this to the rest of the sample as before. Similarly to the previous approach, we would like to fit the prediction model to a set of data where truncation is less problematic. We therefore run specifications where this prediction model is applied to the first 10 years of data (1985–1994).⁹ Table 4 displays results after correcting for truncation. The results of the first, 'fixed' truncation model are presented in the first column, while those from 10-year quadratic model are presented in column 2.

Both methods of correcting for truncation produce similar results—adopting the MMUF does not appear to have a significant effect on an institution's URM PhD completions. Using the same method as before, the largest potential effect size for arts and sciences within a 95% CI of the point estimate is 1.89 PhDs per cohort for the model under the 10-year truncation adjustment. The point estimate itself predicts only 0.518 additional PhDs per cohort—larger than the estimates produced without correcting for truncation, but not statistically significant.

Our analyses so far have used only the MMUF schools, using those institutions in years before they began participating in the program as controls. Although our estimates are fairly precise, we would like to introduce additional control observations. Estimating the program effect using the sample of all U.S. institutions would greatly overstate the effects of the program, as many MMUF institutions were selected for participation in the program specifically because they are high-quality colleges and universities where PhD production is already high. Instead we select two matched control groups constructed using the Stata

⁹ We investigated a similar quadratic model using the first five years of data, and the results were qualitatively similar.

Table 5
Effect of MMUF participation on URM PhD production—matched comparison.

		(1)	(2)
(a)	A&S	0.205 (0.586)	0.127 (0.653)
(b)	A&S + Eng.	0.143 (0.631)	0.131 (0.700)
(c)	All fields	0.301 (0.630)	0.219 (0.726)
	1 Nearest neighbor Kernel	✓	✓

Notes: Marginal effects from six negative binomial models are reported. For each model, the dependent variable is the predicted number of PhD completions among those non-white, non-Asian students who graduated from an institution in a particular year, with degrees in a particular group of fields as indicated by (a), (b), and (c). Prediction is a truncation adjustment with a quadratic model in time fit to the first 10 years of data. All models include the BA completion count for that cohort as well as year and institution fixed effects. Standard errors in parentheses are clustered by institution.

A * indicates significance at the 5% confidence level, and a † the 1% confidence level.

command `psmatch2` to estimate the probability that each institution would be selected to participate in the MMUF program based on its observable characteristics.¹⁰ The first control group uses 1-nearest-neighbor matching to select the non-treated institution with the nearest propensity score to each treated institution as an appropriate control.¹¹ We match with replacement, meaning that a non-treated school can serve as the match for more than one treated school if no other non-treated school is a ‘better’ control.

The second matched control group is constructed using kernel matching to construct an appropriate control institution from a combination of non-participating schools. A list of the variables employed in the matching routine appears as [Appendix Table A3](#), and a list of the schools in the 1-nearest neighbor match appears as [Appendix Table A4](#).

Results from the matching procedures are presented in [Table 5](#). These estimates employ the 10-year flexible truncation correction described above. We find no evidence that participation affected the PhD completion count in arts and sciences, and in fact our point estimates and their standard errors are very close to those in [Table 4](#) that do not include a matched comparison group.

4.2. Estimates of program intensity

In addition to changes in whether a school was participating in the MMUF in a given year there is considerable variation in the size of a MMUF cohort for a given school. This variation should improve our ability to identify the effect of the program and offer an estimate of the effect of changing the size of the program at an institution.

¹⁰ The CUNY schools are excluded from this approach, as their information on the predictor variables we use is reported at the system level and thus we cannot separate treated from non-treated CUNY schools.

¹¹ We estimated the same models with Mahalanobis-metric matching and obtained qualitatively similar results.

Table 6
Effect of intensity of MMUF participation on URM PhD production.

		(1)	(2)
(a)	A&S	0.171 (0.089)	−0.582 (2.567)
(b)	A&S + Eng.	0.188* (0.094)	0.961 (2.646)
(c)	All fields	0.178 (0.096)	−0.947 (4.019)
	Count Dosage	✓	✓

Notes: Marginal effects from six negative binomial models are reported. For each model, the dependent variable is the predicted number of PhD completions among those non-white, non-Asian students who graduated from an institution in a particular year, with degrees in a particular group of fields as indicated by (a), (b), and (c). Prediction is a truncation adjustment with a quadratic model in time fit to the first 10 years of data. All models include the BA completion count for that cohort as well as year and institution fixed effects. Column (1) presents estimates of the marginal effect of adding an additional fellow on the number of PhDs completed. Column (2) presents estimates of the marginal effect of increasing the program size as a proportion of the non-white, non-Asian graduating class. Standard errors in parentheses are clustered by institution.

A * indicates significance at the 5% confidence level, and a † the 1% confidence level.

We do this first by estimating the effect of increasing the number of fellows ($MMUFellows_{it}$) on the number of PhDs completed by URM students ($PhDsM_{it}$). As in the baseline model we include year fixed effects (T_t) and institution fixed effects (I_i). We include a control for the number of BAs completed by URM students ($BAAsM_{it}$). Once again we estimate using negative binomial and regression so the mean of $PhDsM_{it}$ is given by:

$$E(PhDsM_{it}) = I_i \exp(\beta_1 MMUFellows_{it} + \beta_2 BAAsM_{it} + T_t) \quad (3)$$

We find uniformly positive point estimates for the effect of adding an additional fellow, but these estimates are not statistically significant at conventional levels ([Table 6](#)). If correct, our estimates would imply that each student added to the MMUF program adds about 0.171 arts and sciences PhDs that otherwise would not have been completed. This null finding is particularly interesting as the estimates include both the effect of adding a fellow and whatever factors drove a school to add that fellow, which likely includes the strength of a given cohort.

As an alternative to estimating the effect of adding a given number of fellows, we estimate the following, where $Dosage_{it}$ is defined as the number of fellows from graduation cohort t at institution i divided by the number of URM students in that cohort and institution. This model gives the expectation of $PhDsM_{it}$ as:

$$E(PhDsM_{it}) = I_i \exp(\beta_1 Dosage_{it} + \beta_2 BAAsM_{it} + T_t) \quad (4)$$

Increasing the dosage of the program appears to decrease the PhD completion count for all field groups but arts and sciences plus engineering, although this association is nowhere statistically significant. Once again, using a 95% confidence interval around the negative binomial model's estimate for arts and sciences we could rule out a change in expected arts and sciences PhD completions of more than 4.45 per cohort from a 100% increase in the

Table 7
Event study of the effect of MMUF adoption on the URM PhD completion rate.

	t-5	t-4	t-3	t-2	t-1	t	t+1	t+2	t+3	t+4	t+5
Arts and sciences	1.901† (0.633)	-0.343 (0.692)	0.691 (0.647)	-0.128 (0.686)	-0.188 (0.591)	0.759 (0.686)	0.181 (0.610)	-0.185 (0.489)	1.151 (0.776)	-0.237 (0.560)	-0.448 (0.600)
Arts sci. and eng.	1.933† (0.748)	-0.060 (0.695)	0.847 (0.666)	-0.169 (0.737)	0.059 (0.658)	0.701 (0.706)	0.071 (0.638)	-0.461 (0.563)	1.391 (1.036)	-0.073 (0.663)	-0.061 (0.719)
All fields	2.035† (0.711)	-0.179 (0.752)	0.856 (0.632)	0.053 (0.660)	0.106 (0.672)	0.743 (0.691)	0.376 (0.664)	-0.481 (0.578)	1.470 (1.104)	0.220 (0.686)	-0.264 (0.758)

Notes: Marginal effects from thirty three negative binomial models are reported. For each model the dependent variable is the predicted number of PhD completions among those non-white, non-Asian students who graduated from an institution in a particular year relative to the first MMUF cohort at that institution (t). Prediction is a truncation adjustment with a quadratic model in time fit to the first 10 years of data. All models include the BA completion count for that cohort as well as year and institution fixed effects. Standard errors in parentheses are clustered by institution. A * indicates significance at the 5% confidence level, and a † the 1% confidence level.

Table 8
Effect of MMUF participation on the URM PhD completion rate – unadjusted model.

		(1)	(2)	(3)	(4)
	Model	OLS	OLS	OLS	OLS
(a)	A&S	0.011 (0.016)	-0.001 (0.005)	0.013 (0.018)	-0.001 (0.005)
(b)	A&S + Eng.	-0.010 (0.009)	-0.003 (0.005)	-0.010 (0.009)	-0.003 (0.004)
(c)	All fields	-0.014 (0.009)	-0.005 (0.004)	-0.014 (0.009)	-0.005 (0.004)
	Weights		✓		✓
	Non-URM PhD completion rate			✓	✓

Notes: OLS coefficients from 12 models are reported. For each model, the dependent variable is the rate of PhD completion among those non-white, non-Asian students who graduated from an institution in a particular year, with degrees in a particular group of fields as indicated by (a), (b), and (c). All models include year and institution fixed effects. Specifications (3) and (4) include the comparable rate for white and Asian students. Weights are by size of institution in number of URM BA completers. Standard errors in parentheses are clustered by institution. A * indicates significance at the 5% confidence level, and a † the 1% confidence level.

program dosage. This would correspond to expanding the MMUF to cover each school's entire URM population, and in that context is quite a small effect.

4.3. Robustness

We explore the possibility that the adoption of the MMUF program is not random by investigating whether there are any changes in PhD completions up to 5 years before and up to 5 years after program adoption. For this analysis we focus on the data adjusted using the 10-year truncation model. We find some evidence of an increase in PhD production for all three degree groups 5 years before the first cohort was eligible to participate (Table 7). However, no trend is apparent. This may suggest that our truncation adjustment is insufficient to account for the issue of degrees in progress, or simply be a data artifact.

Rather than analyzing the number of PhDs completed by URM students, it is possible to instead consider the PhD completion rate—the percentage of URM bachelors' recipients who go on to complete a PhD. This model would be more appropriate than those discussed previously if institutions with larger numbers of URM students completing bachelors' degrees could expect greater gains, as might be the case if the fellowship somehow changed the expectations or attitudes of both non-participants and participants. This strategy also allows us to experiment with weighting observations by the number of URM bachelors' degrees. This simple weighting scheme has the potential

to increase precision, but is only optimal if the effect of the program on completion rates is homogenous across institutions and the likelihood of PhD completion is not correlated within institution cohorts after controlling for institution and year fixed effects (Solon, Haider, & Wooldridge, 2015). Because these assumptions may not be satisfied we also report the unweighted results. Using our baseline specifications, we find no evidence that the MMUF increases PhD completion rates, and are able to rule out increases larger than five percentage points as outside of a 95% confidence interval of any of the baseline models (Table 8). Results from other specifications lead us to similar conclusions as those using the number of PhDs completed, and are presented in Appendix B (Appendix Tables B1–B3).

We also conduct our baseline estimates with institution-specific linear trends to allow for the possibility that each institution follows its own trend. The results are similar to those found without the inclusion of institution-specific time trends, and are not included for brevity.

5. Conclusion

We describe the Mellon Mays Undergraduate Fellowship Program, the supports it offers its participants, and its growth over time. Using a census of undergraduate completions from the Department of Education as well as a census of PhD completions from the National Science

Foundation, we then attempt to estimate the causal effect of the MMUF program on the PhD completions of URM bachelors' graduates at participant schools. We find no statistically significant effect of an institution's participation in the program and a 95% confidence interval rules out an effect of more than about 1 PhD per cohort using our baseline estimates. We also find no significant effect of increasing the number or percentage of fellows, although we do find predominantly positive point estimates for the number of fellows that would suggest an effect of about 0.171 additional PhDs for each additional fellow. In both cases these estimates are small relative to the 25% of MMUF fellows who are expected to complete PhDs—suggesting an *upper bound* of 68% on the proportion of MMUF fellows who complete PhDs who would not have done so otherwise.

Several factors could explain our null findings. First, the program may simply not do much to increase the number of PhDs produced by URM students. If the program selects the brightest and most motivated students it may benefit those who would have already been likely to attend graduate school and earn PhDs even in its absence. This would not necessarily mean that the MMUF is unimportant—the program could increase the quality of the institutions fellows attend for their doctoral studies, improve dissertations produced or job skills gained, speed completion, or improve the financial position of graduates. Any of these effects could increase the number of URM students entering academia, in addition to being beneficial to fellows, but we are not able to capture them in our data. The Survey of Earned Doctorates is collected at the time of PhD completion and thus is limited in its ability to measure most variables pertaining to careers in academia. Second, the small size of the program at each institution might inhibit our ability to discover an overall effect with statistical models: if only a handful of students in each year are MMUF participants, the largest possible effect the program could have on PhD production will similarly be small. Despite our rather precise estimates we may be failing to detect a real, but small, effect of the program. This is less likely given the insignificant effects we find for increases in program intensity, but those are still complicated by the substantial noise of PhD completions by non-fellows. Finally, our truncation correction could simply be incorrect. If the true truncation pattern is not fit or well approximated by any of our models we may fail to find results where any exist. The true distribution of degree completion times is unknowable until all degrees can be observed, so we cannot rule out this possibility.

Our findings are most generalizable to expansions of the MMUF to institutions relatively similar to those that already participate or increases in the size of the program at MMUF institutions. Many MMUF schools are quite unlike the average U.S. institution and were selected in part based on their high PhD-going rates and a perception that many students had the potential and preparation for a career in academia. However, the program has also been implemented at institutions selected more for the diverse populations they serve (e.g. CUNY schools), where overall student preparation may not be as high. It would be interesting to extend our study to the subsample of the

CUNY schools and the state universities that were later introduced to the program, but that shrinks the pool of observations too greatly to draw meaningful conclusions from the data. There are also other programs that are broadly similar to the MMUF (such as the McNair Scholars Program) to which these same empirical methods could be applied.

Despite the caveats listed above, we hope our findings will prove instructive to designers of future policies. If a program aims to maximize its impact on the number of students achieving any particular benchmark it is important not only to design the program to benefit its recipients but also to select those recipients on the margin of the desired outcome.

Appendix A

Table A1
Non-UNCF Mellon Mays institutions participating by 2005.

Institution name	First year of participation
Barnard College	1998
Bowdoin College	1993
Brown University	1994
Bryn Mawr College	1990
California Institute of Technology	1994
Carleton College	1989
Columbia University	1997
Cornell University	1990
CUNY Brooklyn College	1990
CUNY City College	1990
CUNY Hunter College	1990
CUNY Queens College	1990
Dartmouth College	1990
Duke University	1998
Emory University	2001
Harvard University	1990
Haverford College	2001
Macalester College	2001
Oberlin College	1989
Princeton University	1990
Rice University	1994
Smith College	2000
Stanford University	1989
Swarthmore College	1990
University of Chicago	1990
University of Pennsylvania	1990
University of Southern California	1994
Washington University in Saint Louis	1994
Wellesley College	1990
Wesleyan University	1991
Williams College	1990
Yale University	1990

Table A2
Mellon-designated fields.

Fields as of 2000
Anthropology and Archaeology
Area/Cultural/Ethnic/Gender Studies
Art History
Classics
Demography, Geography and Population Studies
Earth/Environmental/Geological Science and Ecology

(continued on next page)

Table A2 (continued)

English
Ethnomusicology
Film, Cinema and Media Studies (theoretical focus)
Foreign Languages and Literatures
Linguistics
History
Literature
Mathematics
Musicology
Philosophy
Oceanographic/Marine/Atmospheric/Planetary Science
Physics and Astronomy
Political Theory
Religion and Theology
Theater (non-performance focus)
2008 Field Additions
Computer Science
Sociology

Table A3

Predictor variables for propensity score matches.

Institution is Public
Fall Enrollment
Ratio of Undergraduates to All Students
Ratio of Female Undergraduate Students to All Undergraduate Students
Ratio of Full-Time Undergraduate Students to All Undergraduate Students
Ratio of Full-Time Female Faculty to All Female Faculty
Ratio of Female Faculty to All Faculty
Ratio of Full-Time Faculty to All Faculty
Ratio of Faculty to All Staff
Avg. 9–10 Month Salary for All Male Faculty
Avg. 9–10 Month Salary for All Female Faculty
Ratio of Undergraduate STEM Completions to All Undergraduate Completions
Ratio of Undergraduate Humanities Completions to All Undergraduate Completions
Tuition and Fees per Student
Endowment Income per Student
Total Revenues per Student
Instruction Expenditure per Student
Academic Support Expenditure per Student
Student Services Expenditure per Student
Total Scholarship Expenditures per Student
Percentage of Student Body that is Black
Percentage of Student Body that is Asian
Percentage of Student Body that is Hispanic

Table A4

Matched control institutions, nearest neighbor match.

Baptist Bible College of Pennsylvania
Carnegie Mellon University
Case Western Reserve University
Davidson College
Georgetown University
Goucher College
Le Moyne-Owen College
Long Island University

(continued on next page)

Table A4 (continued)

North Carolina State University at Raleigh
Northwestern University
Radcliffe College
Saint Basil's College
San Diego State University
San Francisco Conservatory of Music
Seton Hall University
Smith College
Southern University Agricultural and Mechanical College
University of Massachusetts at Amherst
University of Michigan, Ann Arbor

Appendix B**Table B1**

Effect of MMUF participation on the URM PhD completion rate—truncation adjusted.

	(1)	(2)	(3)	(4)
(a) A&S	0.027 (0.031)	0.001 (0.004)	0.007 (0.015)	−0.001 (0.005)
(b) A&S + Eng.	−0.003 (0.007)	−0.001 (0.004)	−0.011 (0.008)	−0.002 (0.004)
(c) All fields	−0.008 (0.006)	−0.004 (0.003)	−0.014 (0.008)	−0.004 (0.003)
Weights		✓		✓
Simple adjustment	✓	✓		
10-year adjustment			✓	✓

Notes: OLS coefficients from 12 models are reported. For each model, the dependent variable is the predicted rate of PhD completion among those non-white, non-Asian students who graduated from an institution in a particular year, with degrees in a particular group of fields as indicated by (a), (b), and (c). All models include the comparable rate for white and Asian students, as well as year and institution fixed effects. The simple adjustment is a truncation adjustment under the assumption that the time to PhD pattern from 1985 to 1989 persists throughout the sample. The 10-year adjustment is a truncation adjustment with a quadratic model in time fit to the first 10 years of data. Weights are by size of institution in number of URM BA completers. Standard errors in parentheses are clustered by institution.

A * indicates significance at the 5% confidence level, and a † the 1% confidence level.

Table B2

Effect of MMUF participation on the URM PhD completion rate, using a matched comparison group.

	(1)	(2)
(a) A&S	−0.009 (0.006)	−0.000 (0.004)
(b) A&S + Eng.	−0.013 (0.007)	−0.003 (0.003)

(continued on next page)

Table B2 (continued)

		(1)	(2)
(c)	All fields	−0.014 (0.007)	−0.004 (0.003)
	1 Nearest neighbor Kernel	✓	✓

Notes: OLS coefficients from six models are reported. For each model, the dependent variable is the predicted rate of PhD completion among those non-white, non-Asian students who graduated from an institution in a particular year, with degrees in a particular group of fields as indicated by (a), (b), and (c). Prediction is a truncation adjustment with a quadratic model in time fit to the first 10 years of data. All models include the comparable rate for white and Asian students, as well as year and institution fixed effects. Standard errors in parentheses are clustered by institution.

A * indicates significance at the 5% confidence level, and a † the 1% confidence level.

Table B3

Effect of intensity of MMUF participation on the URM PhD completion rate.

		(1)	(2)
(a)	A&S	−0.560* (0.254)	−0.253 (0.099)
(b)	A&S + Eng.	−0.166 (0.092)	0.035 (0.103)
(c)	All fields	−0.250† (0.086)	0.127 (0.173)
	Unadjusted 10-year adjustment	✓	✓

Notes: OLS coefficients from six models are reported. For each model the dependent variable is the predicted PhD completion rate among those non-white, non-Asian students who graduated from an institution in a particular year, with degrees in a particular group of fields as indicated by (a), (b), and (c). Prediction is a truncation adjustment with a quadratic model in time fit to the first 10 years of data. All models include the comparable rate for white and Asian students, as well as year and institution fixed effects. Columns (1) and (2) present estimates of the effect of increasing the dosage of the program on PhD completion rates, with Column (2) additionally adjusting for truncation. Standard errors in parentheses are clustered by institution.

A * indicates significance at the 5% confidence level, and a † the 1% confidence level.

References

- Bengochea, Armando I. (2013). "25th Anniversary review of the Mellon Mays Undergraduate Fellowship (MMUF)." Unpublished manuscript.
- Eagan, M. K., Hurtado, S., Chang, M., Garcia, G., Herrera, F. A., & Garibay, J. (2013). Making a difference in science education: The impact of undergraduate research programs. *American Education Research Journal*, 50(4), 683–713. doi:10.3102/0002831213482038.
- Fairlie, R., Hoffmann, F., & Oreopoulos, P. (2011). A community college instructor like me: Race and ethnic interactions in the classroom. NBER working paper 17381. doi:10.3386/w17381.
- Mellon Foundation. (2003). *The Andrew W. Mellon Foundation: Report from January 1, 2003 through December 31, 2003*. New York: Mellon Foundation Available at http://www.mellon.org/news_publications/annual-reports-essays/annual-reports/content2003.pdf Accessed 24.04.14 .
- Mellon Mays Undergraduate Fellowship (2013). About. Available at www.mmuf.org/about. Accessed 28.04.14.
- National Science Foundation (2012). US citizen and permanent resident doctorate recipients, by race/ethnicity and broad field of study: Selected years, 1991–2011. Available at http://www.nsf.gov/statistics/sed/2011/data_table.cfm. Accessed 24.04.14.
- National Science Foundation (2013). Survey of earned doctorates. Available at <http://www.nsf.gov/statistics/srvydoctorates/#sd>. Accessed 24.04.14.
- Price, J. (2010). The effects of instructor race and gender on student persistence in STEM fields. *Economics of Education Review*, 29, 901–910. doi:10.1016/j.econedurev.2010.07.009.
- Rose, B. (2012). *Program review: The Mellon Mays Undergraduate Fellowship (MMUF) Program*. Wellesley, MA: Brad Rose Consulting Available at <http://bradroseconsulting.com/wp-content/uploads/2012/10/MMUF-Program-Review.pdf> Accessed 28.04.14 .
- Solon, G., Haider, S., & Wooldridge, J. (2015). What are we waiting for? *Journal of Human Resources*, 50(2), 301–316. doi:10.3368/jhr.50.2.301.
- United States Census Bureau (2011). Annual estimates of the resident population by sex, race, and Hispanic origin for the United States: April 1 2010 to July 1, 2011 (NC-ECT2011-03). Available at <http://www.census.gov/popest/data/national/asrh/2011/tables/NC-EST2011-03.xls>. Accessed 08.06.15.