Essays on the Organization of Science and Education

by

Danielle Li

A.B., Harvard College (2005)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Economics
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2012

Author ....
Department of Economics
May 15, 2012

Certified by ....
David H. Autor
Professor of Economics
Thesis Supervisor

Certified by ....
Pierre Azoulay
Associate Professor of Strategy
Thesis Supervisor

Accepted by ....
Michael Greenstone
3M Professor of Environmental Economics
Chairman, Department Committee on Graduate Theses

Essays on the Organization of Science and Education

by

Danielle Li

Submitted to the Department of Economics
on May 15, 2012, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

This dissertation consists of four chapters exploring how organizations inform and distort the implementation of public policy in two empirical settings. Chapters 1 and 2 study the non-market allocation of research funding to scientists while Chapters 3 and 4 examine the market for schools and school leaders.

Experts are likely to have more information regarding the potential of projects in their area, but are also more likely to be biased. Chapter 1 develops a theoretical and statistical framework for understanding and separately identifying the effects of bias and information on expert evaluation and applies it in the context of peer review at the National Institutes of Health (NIH). I use exogenous variation in review committee composition to examine how relationships between reviewers and applicants, as measured by citations, affect the allocation and efficiency of grant funding. I show that, due to bias, each additional related reviewer increases the chances that an applicant is funded by 2.9 percent. Reviewers, however, are also more informed about the quality of proposals from related applicants: the correlation between scores and quality is approximately 30 percent higher for related applicants. On net, the presence of related reviewers improves the quality of research that the NIH supports by two to three percent, implying that reductions in conflicts of interest may come at the direct cost of reducing the quality of funding decisions.

In Chapter 2, I examine how women are treated in grant review at the US National Institutes of Health (NIH). Analyzing funded R01 grants, I show that women receive a half-percentile worse score than men for research that produces the same number of publications and citations. Allowing reviewers to observe applicant gender reduces the number of women who are funded by approximately 3 percent. Analysis of study sections shows that the presence of women attenuates bias, suggesting that diversity in study sections can improve peer review.

Chapter 3 considers the effect of labor market for school leaders. School accountability may affect the career risks that school leaders face without providing commensurate changes in pay. Since effective school leaders likely have significant scope in choosing where to work, these uncompensated risks may limit the ability of low-performing schools to attract and retain effective leaders. This paper analyzes the effect of No Child Left Behind (NCLB) on principal mobility and the distribution of high-performing principals across low- and high-performing schools. I show that NCLB decreases average principal quality at schools serving disadvantaged students by inducing more able principals to move to schools less likely to face NCLB sanctions.

Finally, Chapter 4 explores the viability of voucher base school market reforms by estimating the demand elasticity for private schooling using variation from sibling discounts at Catholic schools. Because families differ in their number and spacing of children, this variation allows us to isolate within-neighborhood variation in tuition prices. We find that a standard deviation decrease in tuition prices increases the probability that a family will send its children to private school by one half percentage point, which translates into an elasticity of Catholic school attendance with respect to tuition costs of -0.19. Our subgroup results suggest that a voucher program would disproportionately induce into private schools those who, along observable dimensions, are unlike those who currently attend private school.

Thesis Supervisor: David H. Autor
Title: Professor of Economics

Thesis Supervisor: Pierre Azoulay
Title: Associate Professor of Strategy

Acknowledgments

It has been a privilege to research and write this dissertation and I am grateful for the support of many individuals and institutions. First and foremost, I owe an enormous debt to my committee chair, David Autor. Beyond bringing his insight to challenging and improving countless iterations of my research, David has consistently been a role model for the type of academic citizen I want to become: one who engages with and values the work of others, takes seriously the task of teaching, cares deeply about learning, and does so with humor and humility. This dissertation would not exist without the enthusiasm and extraordinary efforts of my adviser and co-author, Pierre Azoulay, who helped me access the NIH data that forms the foundation of my work and who has been unfailingly generous with his experience and time. Pierre, more than anyone else, articulated to me a clear, disciplined, and exciting trajectory for the economics of ideas and innovation, to which I would be thrilled to contribute a small fraction of what he has. Michael Greenstone, my final adviser, pushed me to think, act, and communicate with concision and clarity, but showed me incredible patience even when I fell short. Thank you.

In addition, I also thank Daron Acemoglu, Josh Angrist, Esther Duflo, and Bob Gibbons for reading my work and providing me with detailed and constructive feedback. I have greatly benefitted from many discussions with other faculty and fellow students. In particular, Jason Abaluck, Leila Agha, Sue Dynarski, Amy Finklestein, Brigham Frandsen, Jonanthan Gruber, Daniel Keniston, Ziad Obermeyer, Christopher Palmer, Michael Powell, Simone Schaner, Pian Shu, Scott Stern, Xiao Yu Wang, and Heidi Williams deserve special thanks. I would also like to thank participants at the MIT labor and applied micro lunches, the MIT Labor/PF seminar, and the NBER Productivity group for their helpful comments.

I am grateful to Richard Freeman for help in accessing NIH data; Raviv Murciano-Goroff and James Vines for assistance with Web of Science and PubMed; Kara Bonneau and Clara Muschkin for granting me access to data from the North Carolina Education Research Data Center; and Jim Davis for his continual support at the Census Restricted Data Center. Financial support from the MIT Department of Economics and the George and Obie Shultz Fund is gratefully acknowledged.

I have now been fortunate enough to live in Cambridge for eleven years; thank you to Jason Abaluck, David Berger, Michelle Dionne, Soojin Nam, Cindy Paladines, Joshua Reyes, Lisa Shu, and Allison Tanenhaus for making this a happy home, and to Chris Leighton, for having been family.

Finally, I would like to thank my parents, Shiyan Chao and Xiaoxiong Li, who raised me with much love and at great cost. This thesis is dedicated to them.

Contents

Chapter 1

Information, Bias, and Efficiency in Expert Evaluation: Evidence from the NIH

1.1 Introduction

How much should we trust advice from potentially biased experts? Experts may have valuable information about a project's potential, but they may also have preferences that compromise their objectivity. There are many empirical contexts in which these concerns are relevant: corporate boards, venture capital groups, and federal regulatory bodies, for instance, all benefit from the expertise of industry insiders but may also be misled by their advice. This tension between information and bias is especially pronounced when decisions are complex and technical; there is both greater value placed on expertise and greater scope for obfuscation. Particularly in these cases, understanding how to improve the quality of decision-making is difficult because reducing conflicts of interest can come at the direct cost of reducing information.

This chapter develops a framework for separately identifying the effects of bias from that of information and provides the first empirical estimate of the efficiency tradeoff between bias and information in expert evaluation. I do so in a context that is extremely important for medical innovation: grant funding at the National Institutes of Health (NIH). With an annual budget of 30 billion dollars, the NIH is the world's largest funder of biomedical research, spending nearly half as much on basic and applied science as the entire US pharmaceutical industry combined.[1] NIH-sponsored research plays a role in the development of over half of all FDA approved drugs, including path-breaking treatments such as Gleevec, the first drug therapy to selectively target cancerous cells, and Lipitor, one of the most prescribed drugs in America.[2]

The majority of NIH funds are allocated via a non-blind review process in which individual scientists propose research projects that are then evaluated by committees of their peers. Peer review is the key institution responsible for consolidating thousands of investigator-initiated submissions into a concrete, publicly funded research agenda. The success of this system, then, in large part depends on the ability of reviewers to identify and fund the most promising ideas in their areas of specialty.

This chapter evaluates the role that potentially biased reviewers play in NIH peer review. Reviewers may be more qualified to assess the merit of proposals in their own area of expertise, but they may also have conflicts of interest that limit their reliability. I formalize this intuition with a model of strategic communication

[1] In 2006, pharmaceutical companies spent close to 50 billion dollars on R&D. CBO "Research and Development in the Pharmaceuticals Industry" (2006).

[2] Over two-thirds of FDA priority review drugs cite NIH-funded research. See Sampat and Lichtenberg (2011).

in review meetings. In this model, reviewers are biased, meaning that they receive an additional payoff from funding related applicants, independent of that applicant's quality. Reviewers, however, may also improve the quality of funding decisions by introducing better information about the quality of proposals from these related applicants. In equilibrium, a grant proposal's likelihood of being funded can be expressed as a function of its quality, the relatedness of the applicant to the committee, and their interaction. The effect of reviewer bias on funding decisions comes through the level effect of relatedness while the effect of better information comes through the interaction effect.

The intuition behind this result is simple and underlies my empirical work: if committees use reviewer-applicant relationships to make inferences about quality, then the effect of being related to a reviewer should be different for high and low quality applicants. In particular, high-quality applicants should benefit from being related to reviewers while low-quality applicants should be hurt. Reviewers are biased, on the other hand, if they are systematically more (or less) likely to fund related applicants regardless of quality.

Peer review at the NIH presents a rare opportunity to get empirical traction on these issues. To do so, I have assembled a new, comprehensive dataset linking almost 100,000 NIH grant applications to the committees in which they were evaluated. I observe many characteristics of the application, including the application's final score, the name of the applicant, demographic information, grant history, and publication history. For each review meeting, I observe the names of all reviewers who attend and the capacity in which they serve. Using names of applicants and reviewers, I create measures of a reviewer's familiarity with an applicant, as measured by whether she has cited him in the past.

In order to separately identify bias and information, I need detailed measures of grant quality and exogenous variation in relatedness. I measure the quality of grant applications by using text-matching algorithms that link grant project titles to titles and abstracts of publications that the grant produces in the future (See Section 1.5 for details). This strategy can be applied consistent for both funded and unfunded grants because the NIH grants I study require applicants to provide very substantial preliminary results. As a result it is standard practice to publish the research outlined in a grant proposal even if the application goes unfunded.

A remaining concern with this approach is that grant funding can still directly affect my measures of application quality. In addition to restricting my measure of quality to publications that are on the same topic as a grant, I also restrict to articles published so soon after grant review that they are unlikely to be directly affected by any grant funds (See Section 1.5.1 and Appendix 1.11 for discussion and robustness tests.)

Finally, remaining measurement error in grant quality can still affect my measures of bias if application quality and relatedness to committee members are correlated. To deal with this, I exploit the institutional structure of review committees to create exogenous variation in relatedness. In particular, the NIH review committees that I study consist of two types of members, "permanent" and "temporary," who have similar qualifications as scientists but substantially different levels of influence in the committee.[3] This distinction allows me to estimate the plausibly causal effect of relationships on committee decisions by comparing decisions for scientists who are related to the same total number of reviewers (which may be endogenous to scientific quality) but who differ in their number of related permanent members.

Together, my measures of quality and exogenous variation in relatedness al-

[3] "Permanent" members are not actually permanent; they serve four-year terms. See Section 1.5.2 for a discussion of permanent versus temporary reviewers.

low me to 1) estimate the effect of being related to a reviewer on an applicant's scores or chances of funding; 2) assess the role of related reviewers both in terms of how they may bias or inform NIH funding decisions; and finally 3) quantify the efficiency consequences of relationships in terms of the quality of research that the NIH supports.

My paper has three primary findings. First, I show that, holding quality constant, every additional permanent member an applicant is related to increases her chances of being funded by 2.9 percent, the equivalent of a one-fifth standard deviation increase in application quality. Second, I show that reviewers shape committee decisions by both increasing bias and improving information. In particular, while bias increases the average likelihood that related applicants are funded, the expertise that reviewers have about related applicants improves the ability to committees to identify high-quality research. I find that the correlation between scores and funding outcomes for applicants related to permanent members is almost 30 percent higher than it is for those who are related to the same number of total reviewers, but to no permanent reviewers. Finally, on net, I show that the gains associated with review by potentially biased experts dominate the losses. Treating related applicants as if they were unrelated—thereby eliminating both bias and information—would reduce the quality of the NIH-supported research portfolio by two to three percent, as measured by future citations and publications. In addition to quantifying the role that bias and information play on average, I also document substantial and persistent variation in how well grant review committees perform. In particular, I show that some of this variation is attributable to differences in how well committees make use of biased experts.

A growing empirical literature contends that management, team practices, and other organizational choices may explain some of the substantial dispersions in productivity that we observe among firms and other entities (Bloom et. al., 2011; Lazear, Shaw, and Stanton, 2011; Bandiera, Barankay, and Rasul, 2009; Garicano and Heaton, 2007; Ichniowski, Shaw, and Prennushi, 1997). This chapter contributes in this spirit by demonstrating that the organization of review committees matters for how well the NIH allocates funding for scientific research.

My empirical setting is of particular relevance for innovation policy. Many studies focus on evaluating the effect of *receiving* public research funds, either in the form of tax credits (e.g. Hall, 1994) or grant programs (e.g. Lerner, 1999). Jacob and Lefgren's 2011 study notably uses similar NIH administrative data to study the effect of receiving an NIH grant on research outcomes and find very modest effects. There has been significantly less work, however, on the complementary question of how these public research dollars are *allocated.*[4] For example, NIH's reliance on peer review of individual grants stands in contrast with major European funding agencies, which often support large groups of scientists and guarantee their salary. Understanding the strengths and weaknesses of these models is of particular importance because, by making investments in specific people, labs, and ideas, funding not only affects near-term scientific output but may also shape the allocation of future research attention and resources.

This chapter also relates to a large literature on statistical and taste-based discrimination (Becker, 1957; Altonji and Pierret, 2001). My model of grant review adds strategic communication as in Crawford and Sobel (1982) to a signal extraction framework similar to that used in Autor and Scarborough (2008). Like Mobius and Rosenblat (2006) and Chandra and Staiger (2010), I use direct measures of performance outcomes to quantify the efficiency consequences of discrimination.

[4] One recent exception is Hegde (2009), which considers the political economy of NIH congressional appropriations.

Finally, my research brings a quantitative perspective to a primarily sociological literature on how talent is identified (see Merton, 1968, on the allocation of credit in science and more recently Lamont, 2010, on subjectivity in academic peer review).

The remainder of this chapter proceeds as follows. In the next section, I discuss the details of NIH grant review. I discuss my conceptual and statistical frameworks in Sections 3.3 and 1.4, respectively. Section 1.5 explains how I construct my dataset and variables in order to identify the role of bias and information. Main results are presented in Section 1.6. Section 1.7 discusses implications for efficiency and the final section concludes.

1.2 Institutional Context

Each year, thousands of scientists travel to Bethesda, Maryland where they read close to 20,000 grant applications and allocate over 20 billion dollars in federal grant funding. During this process, over 80 percent of applicants are rejected even though, for the vast majority of biomedical researchers, winning and renewing NIH grants is crucial for being an independent investigator, maintaining a lab, earning tenure, and paying salaries.

The largest and most established of these grant mechanisms is the R01, a project-based renewable research grant which constitutes half of all NIH grant spending and is the primary funding source for most academic biomedical labs in the United States. There are currently 27,000 outstanding awards, with 4,000 new projects approved each year. The average size of each award is 1.7 million dollars spread over 3 to five years and the application success rate is approximately 20 percent.

At the NIH, applications are assigned to a review committee, called a "study section," for scoring and to an Institute or Center (IC) for funding. Study sections assess the scientific merit of applications by assigning them a "priority score," which, during the period my data come from, ranged from 1.0 for the best application to 5.0 for the worst, in increments of 0.1. Up to three reviewers read the application and present their initial scores. All members then discuss and anonymously vote on the application using the scores of initial reviewers as a guide. The final score is the average of all member scores. This priority score is then converted into a percentile from 1 to 99, where a percentile reflects the percentage of applications from the same study section and reviewed in the same year that received a better priority score. However, for ease of exposition and intuition, I report percentiles to mean the percentage of applications that are worse, so that higher percentiles are better. For more details, see Gerin (2006).

Once an application has been scored, it is funded in order of score by the IC to which it was assigned, until that IC's budget is exhausted. The lowest percentile score that is funded is known as the payline. A grant's score affects its chances of being funded, but not its actual funding amount; NIH will choose to fund one large grant instead of two or three smaller grants as long as the larger grant has a better score, even if it is only marginally better. Scores are never made public.

The bulk of R01 applications are assigned to one of about 180 "chartered" study sections, which are standing review committees organized around a particular theme, for instance "Cellular Signaling and Regulatory Systems" or "Clinical Neuroplasticity and Neurotransmitters." My analysis focuses on these committees. Chartered study sections meet three times a year in accordance with NIH's three annual funding cycles. During each meeting, they review, on average, 40 to 80 grant applications. Chartered study sections are typically comprised of 15 to 30 "permanent" members who are elected to serve four-year terms and 10-20 "temporary" reviewers, who are called in as needed. The division of committees into permanent and temporary members plays an important role in my identification strategy and I discuss this in greater detail in Section 1.5.2.

1.3 How do Relationships Impact Funding Decisions? Conceptual Framework

The following model of decision-making illustrates how the biases and expertise of an individual reviewer may affect grant allocation through strategic communication. In this model, committees want to fund the best grant applications, but must rely on the recommendation of a reviewer who is potentially biased.

Grant applications have some true quality Q^* that is unobserved by the committee, but which can be observed with varying noise by the reviewer. A reviewer is either related or unrelated. A related reviewer forms a posterior $Q_R = Q^* + \varepsilon_R$ about the quality of the grant and an unrelated reviewer forms the posterior $Q_{UR} = Q^* + \varepsilon_{UR}$. I assume that $\text{Var}(\varepsilon_{UR}) > \text{Var}(\varepsilon_R)$, meaning that a related reviewer is more informed about the true quality of the grant. A related reviewer, however, may be biased: if the grant is funded, he receives a payoff $P^R = Q^* + B$, where B is known. Without loss of generality, I assume that $B > 0$. Neither the committee nor the unrelated reviewer are biased; they receive payoffs of $P^C = Q^*$ and $P^{UR} = Q^*$, respectively. If the grant goes unfunded, all parties receive a common outside option U. The committee can observe whether a reviewer is related or unrelated. I assume that the committee acts as a single unit.

The timing works as follows:

1. Nature draws true quality Q^* and the posteriors Q_R and Q_{UR}.

2. The reviewer, knowing her posterior, makes a costless and unverifiable recommendation $M \in \mathbf{M} = \{M_1, \ldots, M_K\}$ to the committee.

3. The committee observes M and takes a decision $D \in \{0, 1\}$ of whether or not to fund the grant.

4. True quality is revealed and the reviewer and committee both receive their payoffs.

Proposition 1.3.1 *The Perfect Bayesian equilibria of this game are given by:*
Case 1: If $R = 0$, then all informative equilibria are payoff-equivalent to a full-revelation equilibrium in which:

1. *The reviewer truthfully reports her posterior $Q^* + \varepsilon_{UR}$.*

2. *The committee funds the grant if $E(Q^*|Q^* + \varepsilon_{UR}) > U$.*

Case 2: If $R = 1$ then:
For $E(Q^*|Q^* + \varepsilon_R > U - B) > U$, the unique informative equilibrium is partially-revealing:

1. *With probability one, the reviewer sends a signal Y if $Q^* + \varepsilon_R > U - B$ and N otherwise.*

2. *The committee funds the grant if and only if it receives the signal Y.*

In all cases where an informative equilibrium exists, there also exist uninformative equilibria where the grant is never funded.
For $E(Q^*|Q^* + \varepsilon_R > U - B) < U$, only uninformative equilibria exist and the grant is never funded.

Proof: See Appendix 3.3.

Reviewers in this equilibrium signal according to their preferences but, as in Crawford and Sobel (1982), information is distorted because the committee is unable to distinguish when an application reviewed by a related reviewer should be funded (e.g. when $Q^* > U$) from some cases when it should not be (e.g. when $U > Q^* > U - B$). In order for an informative equilibrium to exist, however, committees must believe that enough information about the true quality of the grant is communicated in spite of the distortionary impact of bias.

I will focus on the informative equilibrium both in cases when $R = 0$ and in cases when $R = 1$. The equilibrium message strategy is given by:

$$M(Q) = \begin{cases} Y & \text{if } E(Q^*|Q^* + \varepsilon_{UR}) > U \text{ and } R = 0 \\ Y & \text{if } E(Q^*|Q^* + \varepsilon_R) > U - B \text{ and } R = 1 \\ N & \text{otherwise} \end{cases}$$

and the equilibrium decision strategy is given by:

$$D(M) = \begin{cases} Y & \text{if } M = Y \\ N & \text{otherwise} \end{cases}$$

The equilibrium decision rule can be more succinctly expressed as:

$$\begin{aligned} D &= \underbrace{\mathbb{I}(Q^* + \varepsilon_{UR} > U)}_{\text{baseline for unrelated}} + \underbrace{[\mathbb{I}(Q^* + \varepsilon_R > U) - \mathbb{I}(Q^* + \varepsilon_{UR} > U)]}_{\text{additional information for related (+/-)}} R \\ &+ \underbrace{[\mathbb{I}(U > Q^* + \varepsilon_R > U - B)]}_{\text{bias for related (+)}} R \end{aligned} \qquad (1.1)$$

Equation (1.1) shows that committees have some baseline performance that is captured by how well unrelated reviewers assess the quality of a grant. Advice from related reviewers can improve committee decisions because it increases the chances that a qualified related applicant, one with $Q^* > U$, is funded while decreasing the chances that an exceptionally unqualified related applicant, one with $Q^* < U - B$, is funded. Related reviewers, however, can worsen committee performance by increasing the probability that a related applicant with quality between U and $U - B$ is funded.

In this model, committees listen to the advice of related reviewers even if they are biased because committees value expertise. If the equilibrium were not informative, then advice from related reviewers would not be taken; I would find no effect of bias and perhaps a lower correlation between scores and quality for applications reviewed by related reviewers.

1.4 How do Relationships Impact Funding Decisions? Statistical Framework

Next, I assume that the committee decisions I observe are generated by the equilibrium decision rule described by Equation (1.1) in Section 3.3. Under the assumption that ε is uniform ($\varepsilon_{UR} \sim U[-a_{UR}, a_{UR}], \varepsilon_R \sim U[-a_R, a_R]$) the conditional mean of D is given by:

$$\begin{aligned}
E[D|Q^*, R, U] &= \Pr(Q^* + \varepsilon_{UR} > U) + [\Pr(Q^* + \varepsilon_R > U) - \Pr(Q^* + \varepsilon_{UR} > U)]\, R \\
&+ \Pr(U > Q^* + \varepsilon_R > U - B)R \\
&= \frac{1}{2a_{UR}} [a_{UR} - U + Q^*] + \frac{B}{2a_R} R \\
&\quad + \left(\frac{1}{2a_R} [a_R - U + Q^*] - \frac{1}{2a_{UR}} [a_{UR} - U + Q^*] \right) R \\
&= \frac{1}{2} + \underbrace{\frac{1}{2a_{UR}}}_{\text{Quality corr.}} Q^* + \underbrace{\frac{B}{2a_R}}_{\text{Bias term}} R + \underbrace{\left[\frac{1}{2a_R} - \frac{1}{2a_{UR}} \right]}_{\text{Add. corr. for related}} RQ^* \\
&\quad - \frac{U}{2a_{UR}} + \left[\frac{1}{2a_{UR}} - \frac{1}{2a_R} \right] RU
\end{aligned} \tag{1.2}$$

Many critiques of NIH peer review are based on the claim that related applicants may be more likely to get funded than unrelated applicants, even if their proposals are of similar quality. The underlying assumption is that this difference in funding likelihood is due to bias. Equation (1.2) shows, however, that the effect of relationships on the allocation of grant funding is actually more nuanced.

Relationships can increase the likelihood that an application is funded either because of bias or because related reviewers know more about an applicant's quality and can thus increase a committee's confidence in a proposal. The latter effect comes through the RQ^* term—related applicants with high quality will be more likely to be funded. Distinguishing between these cases is important because they have different implications for whether relatedness enhances the quality of peer review. Further, even if relatedness does not affect the likelihood of that an applicant is funded on average, relatedness can still affect the probability that a particular applicant gets funded. If reviewers have more information about the quality of related applicants, then high quality related applicants should be more likely to be funded than high quality unrelated applicants but low quality related applicants should be less likely to be funded than low quality unrelated applicants. Relatedness can thus have a main effect on the likelihood that an applicant is funded and an interaction effect with quality.

Equation (1.2), moreover, says that the effect of relatedness coming from bias and from information can be separately identified. The intuition is simple: if reviewers have more information about the quality of related applicants, then the effect of relatedness on funding likelihood should differ for high and low quality applicants. If committees were influenced by the bias of related reviewers, then related applicants should be more likely to be funded regardless of quality.

This intuition is reflected in the coefficients on R, Q^*, and RQ^*. The coefficient on R captures the effect of reviewer bias and is non-zero if and only if $B \neq 0$. The coefficient on Q^* describes the quality of information received by unrelated reviewers. This term captures, for unrelated applicants, how well committees translate increases in application quality into increases in the likelihood of being funded. A higher coefficient on Q^* means that a committee is good at identifying and funding high-quality research among unrelated applicants. The coefficient on RQ^*, meanwhile, captures the differential effect of relatedness arising from information. The effect of information is larger when the difference between the precisions of related and unrelated beliefs, $\frac{1}{2a_R} - \frac{1}{2a_{UR}}$ is greater.

Finally, the terms U and RU control for the degree of selectivity; when the cutoff U is high, there is little correlation between funding and quality even in the absence of bias or differential information because it is difficult to distinguish quality when all funded applicants are very high quality. In the model, there is no limit to the number of grants that are funded so that relationships can also affect the generosity of committees. The RU term ensures that relationships are not credited for changing the correlation between funding and quality simply by lowering the threshold at which grants are funded. My results are robust to allowing for non-linear effects of relatedness and quality measures. These results are available from the author.

Equation (1.2) has a somewhat surprising feature: it says that, as long as Q^* is perfectly observed, I do not need exogenous variation in relatedness to identify the presence of bias. This is because exogenous variation in relationships matters only if application quality is an omitted variable. If, however, quality is observed, then exogenous variation in relatedness would not be necessary because I would be able to directly control for quality.

In practice, though, I do not observe a grant's true quality Q^*. Instead, I observe a signal of quality $Q = Q^* + v$. Thus, while the model suggests the following equation:

$$S = \alpha_0 + \alpha_1 Q^* + \alpha_2 R + \alpha_3 RQ^* + \alpha_4 U + \alpha_5 RU + X\beta + \varepsilon \tag{1.3}$$

I can only estimate:

$$S = a_0 + a_1 Q + a_2 R + a_3 RQ + a_4 U + a_5 RU + Xb + e. \tag{1.4}$$

where, in both equations, X includes other relevant variables that I can condition on.

Proposition 1.4.1 *Given observed quality $Q = Q^* + v$, the bias parameter α_2 in Equation* (1.3) *is consistently estimated by a_2 in Equation* (1.4) *as long as the following conditions are met:*

1. $Cov(R, Q^*|U, RU, X) = 0$ *and* $Cov(R^2, Q^*|U, RU, X) = 0$

2. $E(v|U, RU, X) = 0$

3. $Cov(v, R|U, RU, X) = 0$

Proof: See Appendix 1.10.

These are my identifying conditions. Condition 1 requires that my measure of relatedness not be correlated with true application quality, conditional on some set of observables. If this were not the case, any mismeasurement in true quality Q^* would bias estimates of α_2 through the correlation between Q^* and my relatedness measure R. Thus, in my study, exogenous variation in relatedness is required only to deal with measurement error.

Condition 2 requires that measurement error be mean zero conditional on observables. Condition 3 says that the extent of measurement error should not depend, conditional on observables, on whether an applicant is related to a reviewer. Together, these conditions are weaker than classical measurement error.

Condition 3 may not be satisfied if related applicants are more likely to be funded and funding itself affects my measure of quality. Suppose, for instance, that two scientists apply for a grant using proposals that are of the same quality. One scientist is related to a reviewer and is funded because of bias. The funding, however, allows her to publish more articles meaning that my measure of quality, future citations, may mistakenly conclude that her proposal was better than the other scientist's to begin with. Mismeasurement of ex ante grant quality makes it *less* likely that I would find an effect of bias.

Another important reason why Condition 3 may not be satisfied is given by the Matthew Effect, a sociological phenomenon wherein credit and citations accrue to established investigators simply because they are established (see Merton, 1986; Azoulay, Stuart, and Wang, 2011). Were this the case, more related applicants would receive more citations regardless of the true quality of their work, meaning that measurement error v would be correlated with relatedness. The Matthew Effect would also make it less likely that I would find an effect of bias; related applicants may get higher scores simply for being established, but this bias would look justified by my measure of quality (which reflects bias in the scientific community at large).

In the next section, I discuss how my sample and variables are constructed in order to disentangle the effect of bias versus the effect of information. I pay particular attention to describing how I define and measure relatedness and quality in order to meet my identifying conditions, described above.

1.5 Data and Empirical Strategy

In order to understand how relatedness affects committee decisions, I have constructed a new dataset describing grant applications, review committee members, and their relationships for almost 100,000 applications evaluated in over 2,000 meetings of 250 chartered study sections. My analytic file combines data from three sources: NIH administrative data for the universe of R01 grant applications, attendance rosters for NIH peer review meetings, and publication databases for life sciences research. Figure 1 summarizes how these data sources fit together and how my variables are constructed from them.

I begin with two primary sources: the NIH IMPAC II database, which contains administrative data on grant applications and a series of study section attendance rosters obtained from NIH's main peer review body, the Center for Scientific Review. The application file contains information on the full name and degrees of the applicant, the title of the grant project, the study section meeting to which it was assigned for evaluation, the score given by the study section, and the funding status of the application. The attendance roster lists the full names of all reviewers who were present at a study section meeting as well as information on whether a reviewer served as a temporary member or as a permanent member. These two files can be linked using meeting-level identifiers available for each grant application. Thus, for my sample grant applicants, I observe the identity of the grant applicant, the identity of all committee members, and the action undertaken by the committee.

Next, I construct detailed measures of applicant demographics, grant history, and prior publications. Using an applicant's first and last name, I construct probabilistic measures of gender and ethnicity (Hispanic, East Asian, or South Asian).[5] I also search my database of grant applications to build a record of an applicant's grant history as measured by how many new and renewal grants the applicant has received in the past, and the number of these grants that the applicant has applied for. This includes data on non-R01 NIH grants such as post-doctoral fellowships and career training grants. To get measures of an applicant's publication history, I use data from Thomson-Reuters Web of Science (WoS) and the National Library of Medicine's PubMed database. From these, I construct information on the number of research articles that an applicant has published in the five years prior to submitting her application, her role in those publications (in the life sciences, this is discernable from author position), and the impact of those publications as measured by citations. In addition to observing total citations, I can also identify a publication as "high impact" by comparing the number of citations it receives with the number of citations received by other life science articles that were published in the same year.

My final sample consists of 93,558 R01 applications from 36,785 distinct inves-

[5] For more details, see Kerr (2008).

tigators over the period 1992-2005. Of these applications, approximately 25 percent are funded and 20 percent are from new investigators, those who have not received an R01 in the past. This sample is derived from the set of grant applications that I can successfully match to meetings of study sections for which I have attendance records, which is about half of all R01 grants reviewed in chartered study sections. Table 1 shows that my sample appears to be comparable to the universe of R01 applications that are evaluated in chartered study sections.

So far, I have discussed how I measure the prior qualifications of an applicant. As Conditions 1-3 of Section 1.4 indicate, however, I also need a direct measure of grant quality and a measure of relatedness that is conditionally independent of quality. I discuss each of these requirements in turn.

1.5.1 Measuring Quality

A major strength of this project lies in my ability to go beyond past applicant characteristics in assessing application quality. Instead, I am able to observe a direct measure of the quality of an application by looking at the publications and citations it produces in the future. Due to the nature of the R01 grant application process, grant applications are likely to produce publications even when the application is not funded. This is because R01s are intended for projects that have demonstrated a substantial likelihood of success, meaning that R01 applicants are required to produce substantial "preliminary results" as a part of their grant application. In practice these stringent requirements mean that preliminary results are often developed fully enough to be published as standalone articles even if the grant application itself goes unfunded. In fact, the bar for preliminary results is so high that the NIH provides a separate grant mechanism, the R21, for pursuing them.

For every grant application I observe, I find articles published by that grant's primary investigator around the time when the grant was reviewed. These publications, and the citations that they generate, form the basis of my measure of grant quality. As discussed in Section 1.4, however, measurement error in the quality of applications poses several challenges. In particular, I need to find a quality measure that is consistent for funded and unfunded grants and not directly affected by funding.

I tackle the first concern by devising a way to link grant applications to their related publications using only information that would exist for both funded and unfunded grants. In particular, this means that I cannot make use of explicit grant acknowledgements because they are available only for funded grants. Instead, I compare the titles and abstracts of an applicant's publications with the title of her grant proposal to determine which publications are related. For instance, if I see a grant application entitled "Traumatic brain injury and marrow stromal cells" reviewed in 2001 and an article by the same investigator entitled "Treatment of traumatic brain injury in female rats with intravenous administration of bone marrow stromal cells," published around this time, I conclude that this publication and its future citations can be used as a measure of the quality of the grant application. Text-matching ensures that I can measure quality using the same procedure for all grant applications.

The second challenge in assessing quality is to make sure that my measure of quality is not directly affected by funding. Grant funding, for instance, can be used to start new experiments related to the proposed project or to subsidize research on unrelated projects. Existing evidence on the effect of grant funding on research outcomes suggests that this effect is likely to be small; using a regression-discontinuity approach, Jacob and Lefgren (2011) find that receiving an R01 increases the number of articles a PI publishes in the next five years by 0.85, from a mean of 14.5. This figure includes all publications by a PI, including ones that may be on a different topic from the original application. Jacob and Lefgren's analysis, however, only documents the effect of grant receipt for marginal applicants. The effect of funding on future publications and citations could be larger elsewhere in the distribution and I take additional precautions to create a measure of quality not affected by funding.

Text-matching limits the set of publications I use to infer application quality to those which are on the same topic as the grant. This reduces the possibility that my measure of application quality is contaminated by unrelated research that the grant is used to subsidize. I address concerns that grant funding might increase the number of publications related to the grant proposal topic by only considering articles published in a short time window surrounding grant review. These articles are likely to be based on research that was already completed or underway at the time the grant application was written. To compute the appropriate window, I consider funding, publication, and research lags. A grant application is typically reviewed four months after it is formally submitted and, on average, another six months elapse before it is officially funded.[6] In addition to this ten month funding lag, publication lags in the life sciences (the time between first submission and publication) typically range from three months to well over a year. Because running experiments, analyzing data, and writing drafts also takes time, it is unlikely that articles published up to two years after a grant's review would have been directly supported by that grant. I also include related publications published one year before a grant is reviewed because these publications likely contribute to the research that is proposed in the application.

Figure 2 confirms that grant applications produce related publications even if they are unfunded. In fact, using my measure of quality described above, I find that

[6] See http://grants.nih.gov/grants/grants_process.htm.

funded and unfunded grants are almost equally represented among the subset of grant applications that generate many citations. Figure 2 also shows, however, that unfunded grants are more likely to produce few citations. This can either mean that unfunded applications are of lower quality and should thus be expected to produce fewer citations or that funding directly improves research output, meaning that I differentially mismeasure quality for funded and unfunded grants.

I distinguish between these explanations by using year-to-year and subject-to-subject variation in whether grant applications with the same score are funded. If funding has a direct impact on my quality measure, then funded grants should produce more citations than unfunded grants conditional on having the same score. Figure 3 shows that this is not the case. Each dot represents the mean number of citations associated with grant applications that received a particular percentile score, regression adjusted to account for differences across fields and years. The dots represent outcomes and scores for funded grants, the crosses for unfunded grants. The dots and crosses overlap because budgets vary across time and across fields, meaning that similarly ranked grants are sometimes funded and sometimes not. In these areas, outcomes for funded and unfunded grants with the same score are similar. There is no evidence that funding directly improves outcomes.

The accompanying statistical test is reported in Table 2. I compare measured quality for funded and unfunded grant applications with similar scores from applicants with similar characteristics. Funding status can vary if some grants are funded out of scoring order or it can vary because funding varies across fields and years. Columns 1 and 2 show that awarded grants tend to be higher quality, but this effect goes away once I control for a smooth function of scores. Together with Figure 3, this finding mitigates concerns that my measure of quality is directly affected by funding.

I discuss several more robustness tests in Appendix 1.11. First, I show that my results hold if I restrict publications associated with grants to those published one year before and one year after grant review. This short time window means that it would be highly unlikely that an article could be directly supported by grant funding because funding and publication lags themselves are likely to total over a year. Appendix 1.11 also reports another test of the validity of my quality measure. If my results were driven by changes in measured grant quality near the payline, then I should find no effects of relatedness on scores for the subset of grant applications that are either well above or well below the payline. However, in both samples, I do find evidence that being related to a permanent member increases scores and increases the correlation between scores and quality. Because relatedness cannot affect actual funding status in this subset, the effect I find cannot be driven by differences in how well quality is measured.

It is also worth emphasizing that, as discussed in Section 1.4, overcrediting funded applications relative to unfunded applications would lead me to *underestimate* the extent of bias.

1.5.2 Identifying Relationships

Next, I determine whether an applicant and a reviewer are related using their citation history. Specifically, using data from Web of Science, I define an applicant to be related to a reviewer if the reviewer has cited the applicant in the five years prior to the review meeting. Citation relationships capture the extent to which reviewers are aware of an applicant's prior work and whether they find that work useful for their own research. In particular, I assume that reviewers are more likely to be familiar with the work or subfield of authors they cite than authors they do not cite.

Table 3 describes applicant-reviewer relationships in my sample study sections. In total, I observe 18,916 unique reviewers. On average, 30 reviewers attend each meeting, 17 of whom are permanent and 13 of whom are temporary. The average applicant has been cited by two reviewers, one temporary and one permanent. The average permanent and average temporary reviewer both cite four applicants. This relatively low amount of relatedness indicates that citations are capturing a finer measure of expertise than simple field overlap. Because the review committees I study are highly focused, most reviewers will be in the same broad fields as measured by their departmental affiliations—molecular biology, surgery, etc.—citations allow me to get a finer measure of the type of work that reviewers are familiar with and thus more variation in relatedness. Appendix 1.11 discusses robustness to alternative measures of relatedness using mutual citations or restricting citation linkages to publications in which both the reviewer and applicant were primary (first, second, or last) authors.

Whether an applicant has been cited by a reviewer is likely to be correlated with the applicant's quality. Applicants who are prominent scientists may be more likely to be cited by reviewers and they may also be more likely to receive higher scores. This correlation would violate Condition 1 of Section 1.4. I exploit the structure of chartered NIH study sections in order to find exogenous variation in reviewer-applicant relatedness. As discussed in Section 1.2, the review committees I study consist of "permanent" and "temporary" members. Permanent members and temporary members are comparable as scientists. Figure 4 and Table 4 show that they have similar publication histories and demographics. In fact, Table 4 indicates that they are often the same people; 35 percent of current permanent members will work as temporary members in the future and 40 percent of current temporary members will work as permanent members in the future.

Permanent members are also likely to have more influence in general. Because they serve as reviewers for more grants, permanent members exert greater influence over committee decisions by providing more initial scores. Temporary members, moreover, vote on fewer proposals because they are often not expected to stay for meeting days in which their assigned grants are not up for discussion. Finally, permanent members work with each other over the course of four years or twelve committee meetings and are more likely to have relationships with each other. A test of the assumption that permanent members have more influence is reported in Appendix 1.11.

Given this, I identify the effect of relationships by examining how the number of permanent members an applicant is related to, call this R^P, affects the committee decision, conditional on the *total* number of a related reviewers, R. My identification compares the outcomes of scientists whose applications are reviewed in the same meeting, who have similar past performance, who are related to the same total number of reviewers, but who are related to different numbers of permanent reviewers.

Using relatedness to permanent members also addresses concerns about the Matthew Effect. Because my identification holds scientific esteem as measured by total relationships constant, there is no reason to believe that applicants related to permanent members would be more or less likely to be cited than applicants related to temporary members.

Figure 5 provides general evidence that the number of permanent members an applicant is related to is not correlated with her quality, conditional on total relatedness. The first panel shows the distribution of application quality as measured by future citations for applicants related to exactly one reviewer. The solid line shows the distribution for applicants related to one permanent member; the dotted line shows the distribution for those related to one temporary member. These distributions are essentially identical. Similarly, Panel 2 shows that the distribution of application quality is the same whether an applicant is related to two temporary, two permanent, or one temporary and one permanent member.

1.5.3 Estimating Equations

Taking these specific measures of quality and relatedness, my schematic regression from Section 1.4 translates into the following set of estimating equations.

First, using variation in relatedness to permanent members, the effect of relatedness on an applicant's likelihood of funding can be estimated from the following regression:

$$D_{icmt} = a_0 + a_1 R^P_{icmt} + a_2 R_{icmt} + \mu X_{icmt} + \delta_{cmt} + e_{icmt}. \tag{1.5}$$

D_{icmt} is a variable describing the decision (either the score or the funding status) given to applicant i whose proposal is evaluated by committee c in meeting m of year t. R^P_{icmt} is the number of permanent reviewers an applicant is related to and R_{icmt} is the total number. The covariates X_{icmt} include indicators for sex and whether an applicant's name is Hispanic, East Asian, or South Asian, quartics in an applicant's total number of citations and publications over the past five years, indicators for whether an applicant has an M.D. and/or a Ph.D., and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to. The δ_{cmt} are fixed effects for each committee-meeting so that my analysis compares outcomes for grants that are reviewed by the same reviewers in the same meeting. Standard errors are clustered at the committee-fiscal year level. Given these controls, a_1 captures the effect of being related to an additional permanent reviewer on the likelihood that an applicant is funded.

The full effect of relationships on funding decisions, however, is more nuanced. The model in Section 3.3 predicts that both the level likelihood of funding and its slope with respect to quality will be higher for related applicants. To test these predictions, I estimate:

$$\begin{aligned} D_{icmt} &= a_0 + a_1 R^P_{icmt} + a_2 Q_{icmt} \times R^P_{icmt} + a_3 Q_{icmt} \\ &+ a_4 R_{icmt} + a_5 R_{icmt} \times Q_{icmt} + \mu X_{icmt} + \delta_{cmt} + \varepsilon_{icmt} \end{aligned} \quad (1.6)$$

Equation (1.6) uses the same controls as in Equation (1.5) and adds several variables describing the quality of the grant application. $Q_{icmt} \times R^P_{icmt}$ is the interaction between number of permanent reviewers and quality and Q_{icmt} is the level effect of quality on the committee decision D_{icmt}. Equation (1.6) includes a control for the total number of related reviewers interacted with quality, $R_{icmt} \times Q_{icmt}$. This is necessary because the total number of reviewers who cite an applicant may be correlated with an applicant's quality; without this control, the variable of interest $R^P_{icmt} \times Q_{icmt}$ may simply be capturing the difference in correlation between quality Q_{icmt} and committee decisions D_{icmt} for high quality applicants (those cited by more reviewers). For instance, the correlation between scores and quality for well-cited candidates may be mechanically lower than for poorly-cited candidates because it may simply be harder to distinguish among high quality applications. Controlling for $R_{icmt} \times Q_{icmt}$ accounts for this possibility.

In Equation (1.6), the coefficient a_1 is the effect of being related to an additional permanent member on funding that is attributable to bias. The coefficient a_2 measures the information effect of being related to a permanent member. Comparing two scientists related to the same total number of reviewers, a_2 captures the additional change in the likelihood of funding for the applicant related to a permanent member, for the same one unit increase in quality. Equation (1.6) says that if committees are using relationships to make better inferences about the quality of an application, then the effect of relationships should be captured by the interaction of quality and relatedness, $Q_{icmt} \times R^{P}_{icmt}$. Any remaining level effect of relationships is then attributable to bias.

1.6 Main Results

Table 5 considers the effect of being related to a committee member on scores and funding. The first column reports the raw within-meeting association between the number of permanent related reviewers and an applicant's likelihood of being funded. Without controls, each additional related permanent member is associated with a 3.3 percentage point increase in the probability of funding, which translates into a 15.3 percent increase from an average funding probability of 21.4 percent. Most of this correlation, however, reflects differences in the quality of applications; applicants may be more highly cited by reviewers simply because they are better scientists. Column 2 adds controls for applicant characteristics such as past publication and grant history. This reduces the effect of an additional permanent related reviewer on funding probability to 1.5 percentage points or 7.1 percent. Even with these controls, relatedness may still be proxying for some unobserved aspect of application quality. Finally, I control for the total number of reviewers each applicant has been cited by. Given this, my identification comes from variation in the composition of an applicant's related reviewers; I am comparing outcomes for two scientists with similar observables, who are cited by the same total number of reviewers, but different numbers of influential reviewers. In Column 3, I find that an additional permanent related reviewer increases an applicant's chances of being funded by 0.6

percentage points or 2.9 percent. This is my preferred specification because it isolates variation in relatedness that is plausibly independent of an application's quality. I find similar effects when an applicant's score is the dependent variable.

The estimates in Table 5 do not distinguish between the impact of bias and the impact of information. Table 6 reports my main regressions, decomposing these effects. Column 1 and 3 reproduce the estimates of the level effect of relatedness on funding and scores from Table 5. Column 2 reports estimates of the coefficients from Equation (1.6). I show that each additional applicant still increases the likelihood that a grant is funded by 0.6 percentage points or 2.9 percent. Since I also include controls for an application's quality and its quality interacted with relatedness, this figures means that the entire level effect of relationships on funding is likely due to bias.

Column 2 also shows that the review committee does a better job of discerning quality when an applicant is related to a permanent member, conditional on the total number of related reviewers. To see this, consider an applicant who is related to one permanent member versus an applicant who is related to one temporary member. A one standard deviation increase in quality for the former applicant increases her likelihood of funding by 1.06+3.15-0.16 = 4.05 percentage points or 4.05/21.4 = 18.9 percent compared with 3.14-0.16 = 2.99 percentage points or 2.99/21.2=14.0 percent for the latter applicant. Being related to a permanent member, then, increases the ability of the committee to predict application quality by over 30 percent. Thus, despite overall positive bias in favor of related applicants, being related to a permanent member may not be beneficial for all applicants. Because reviewers have more information about the quality of related applicants, related applicants with lower quality proposals end up receiving lower scores. These results are consistent with the predictions of my model: relationships decrease the variance of the committee's signal of quality but also increase the distortion arising from bias.

Column 2 also reports the increase in funding likelihood associated with an increase application quality. The figure of 0.0315 means that a one standard deviation increase in application quality is associated with a 3.2 percentage point or 3.2/21.4=14.9 percent increase funding probability for applicants who are not related to any reviewers at all. The sensitivity of committees to changes in application quality highlights the magnitude of the bias effects that I find: being related to an additional permanent reviewer increases an applicant's chances of being funded by as much as a one-fifth standard deviation increase in quality.

The coefficient on total related reviewers interacted with quality is estimated to be negative. This means that the correlation between quality and funding is lower for applicants related to more reviewers. If total related reviewers were proxying for quality, this result would not be unexpected; it may be harder to distinguish quality among grant proposals from high quality scientists than from low quality scientists, where the variance in quality may be higher overall.

Finally, looking at Column 4, a similar though noisier pattern can be seen for scores. While being related to a reviewer increases the level score that one receives for reasons due to bias, it also improves the correlation between an application's quality and its chances of being funded.

In Table 7, I consider how the role of relationships may differ for new and experienced investigators and for new and competing renewal applications. Approximately 20 percent of grant applications are submitted by scientists who have no prior R01 grants. Understanding how applications from new investigators are treated is of particular importance for identifying and supporting promising young scientists.

Even though they are applying for their first R01 grant, new investigators are not entirely unknown to study sections. Forty percent of would-be new investigators have been cited by a reviewer in the past, indicating that the reviewer may be familiar with their work, or at least the work coming out of their lab. Columns 1 and 2 show that there appears to be little bias in the evaluation of new investigators. Related reviewers also do not have better information about the quality of new investigators even though they do appear to be more informed about the quality of experienced investigators. In fact, the entire effect of bias and information estimated in Table 6 appears to be driven by the evaluation of experienced investigators.

Columns 5-8 of Table 7 consider the effect of relatedness for new versus competing renewal applications. I find that related reviewers have fewer insights or biases about the quality of new grants. In both the case of new investigators and new proposals, the bias and information effects of relationships I find are driven by the subset of grants for which there may already be more information. Because there are substantially more experienced investigators but substantially fewer renewal grants in my sample, this effect is not driven by larger sample sizes or more precise estimates.

1.7 How Do Relationships Affect the Efficiency of Grant Provision?

My main results show that relationships affect committee decisions by increasing bias and increasing information. In this section, I embed my analysis of the effect of relationships on *decisions* into a broader analysis of its effect on *efficiency*. In particular, I estimate the net effect of relationships on the quality of decision-making, assuming that policy makers care about maximizing the number of publications and citations associated with NIH-funded research.

I begin by comparing the actual funding decision for an application to the counterfactual funding decision that would have obtained in the absence of relationships. Specifically, I define:

$$\begin{aligned} D^{\text{Benchmark}}_{icmt} &= D_{icmt} \text{ (actual funding)} \\ D^{\text{No Relationship}}_{icmt} &= D_{icmt} - \widehat{a_1} R^P_{icmt} + \widehat{a_2} Q_{icmt} \times R^P_{icmt} \end{aligned}$$

where $\widehat{a_1}$ and $\widehat{a_2}$ are estimated from Equation (1.6) of Section 1.5.3.[7] The counterfactual funding decision represents what the committee would have chosen had applicants related to permanent members been treated as if they were unrelated.

I summarize the effect of relationships by comparing the quality of the proposals that would have been funded had relationships not been taken into account with the quality of those that actually get funded. Specifically, I consider all applications that are funded and sum up the number of publications and citations that accrue to this portfolio. This is my benchmark measure of the quality of NIH peer review. I then simulate what applications would have been funded were relationships not taken into account. To do this, I fix the total number of proposals that are funded in each committee meeting but reorder applications by their counterfactual funding probabilities. I sum up the number of publications and citations that accrue to this new portfolio of funded grants. The difference in the quality of the benchmark and counterfactual portfolio provides a concrete, summary measure of the effect of relationships on the quality of research that the NIH supports.

To get a fuller sense of how committees affect decision-making, I create a measure of committee-specific performance and examine how relationships affect the distribution of performance among NIH peer review committees. First, I define a

[7] Even though $D^{\text{No Relationship}}_{icmt}$ is constructed using estimates from Equation (1.6), it does not rely on the model to interpret those coefficients.

committee's *value-added.* Suppose two scientists submit applications to the same committee meeting. A good committee is one that systematically funds the application that is higher quality. Good committees, moreover, should bring insights beyond what can simply be predicted by objective measures of an applicant's past performance. In particular, suppose now that two scientists with identical objective qualifications submit applications to the same committee meeting. A committee with high value-added is one that systematically funds the application that subsequently generates more citations, even though the applications initially look similar. My measure of committee value-added formalizes this intuition:

$$D_{icmt} = a + b_{cmt}Q_{icmt} + \mu X_{icmt} + \delta_{cmt} + e_{icmt}. \tag{1.7}$$

Here, the dependent variable is either an application's actual funding status $D_{icmt} = D_{icmt}^{\text{Benchmark}}$ or its counterfactual funding status $D_{icmt} = D_{icmt}^{\text{No Relationship}}$. The committee fixed effects δ_{cmt} restrict comparisons of applications to those evaluated in a single meeting and the X_{icmt} control for past applicant qualifications. The coefficients of interest are the b_{cmt}. These are meeting specific slopes that capture the relationship between an application's quality Q_{icmt} and its likelihood of being funded D_{icmt}. Each b_{cmt} is interpreted as the percentage point change in the likelihood that an application is funded for a one unit increase in quality. This forms the basis of my committee value-added measure.

This concept of committee value-added differs from the classical notion of value-added commonly used in the teacher or manager performance literature (see Kane, Rockoff, and Staiger 2007, and Bertrand and Schoar 2003). Teacher value-added, for instance, is typically estimated by regressing student test scores on lags of test scores, school fixed effects, and teacher fixed effects. A teacher's xed effect, the average performance of her students purged of individual, parental, and school-wide inputs, is taken to be the basic measure of quality.

This traditional measure, however, does not capture value-added in my setting. Good committees are not ones in which all applications are high performing; after all, committees have no control over what applications get submitted. Rather, good committees are ones in which funded grants perform better than unfunded grants. I measure a committee's performance by the relationship between an applicant's quality and its likelihood of getting funded because, unlike a teacher, a committee's job is not to *improve* the quality of grant applications but to *distinguish* between them.

One concern with the estimated $\hat{b}_{cmt}$ is that idiosyncratic variation in grant performance may lead me to conclude that some committee meetings do an excellent job of identifying high quality applications when in fact they are simply lucky. I correct for this by modeling $\hat{b}_{cmt}$ as a combination of the committee's true value-added plus a noise term, which I assume to be independent and normal:

$$\hat{b}_{cmt} = b^*_{cmt} + \nu_{cmt} \tag{1.8}$$

Using an empirical Bayes estimator, I adjust $\hat{b}_{cmt}$ for sampling variation so that I define committee quality based only on the portion of $\hat{b}_{cmt}$ that is correlated across multiple meetings; an estimate $\hat{b}_{cmt}$ is taken seriously only if it is consistent across multiple meetings of that committee within the same fiscal year. Otherwise, the Bayesian shrinkage estimator reweights that observation toward the mean. Appendix 1.12 describes this procedure in more detail.

1.7.1 Results

Table 8 estimates the effect of relationships on the quality of research that the NIH supports. In effect, I ask what the NIH portfolio of funded grants would have been had committees treated applicants who are related to permanent members as if they were not, holding all else fixed. In my sample, I observe 93,558 applications, 24,404 of which are funded. Using this strategy, I find that 2,166 or 2.3 percent of these applications change funding status under the counterfactual.

On average, relationships help applicants get funded so that ignoring them would decrease the number of related applicants who are funded by 3.5 percent. These applications from related reviewers, however, are on average better than the applications that would have been funded had relationships not mattered. The overall portfolio of funded grants under the counterfactual produces two to three percent fewer citations, publications, and high impact publications.

This pattern is underscored by Figure 6, which graphs the distribution of value-added under the benchmark and counterfactual cases. Under the benchmark, a one standard deviation increase in the quality of an application evaluated by the median committee would increase its likelihood of funding by approximately 14.5 percent. When relationships are ignored, this figure falls to 11.1 percent.

Figure 6 also shows that there is significant variation in the ability of committees to identify grant applications that subsequently produce high-impact research. Regardless of whether relationships inform committee decisions, the bottom quarter to third of committees actively subtract value, meaning that increases in quality are correlated with *decreases* in the likelihood that an application is funded. As explained in Section 1.7, these figures account for sampling variation so that a committee is deemed to have negative value-added only if it systematically does so from meeting to meeting.

Table 9 presents preliminary evidence that good committees are able to make better use of expert information while limiting the extent of bias. In this table, I run my main regressions on separate samples of high and low performing committees according to the value-added measure discussed in Section 1.7. All results are weighted by the precision of the value-added estimate. Columns 1 and 2 present the main results on bias and information estimated separately for above and below median committee meetings. Although the standard errors are large, relationships appear to affect the decisions of below median committees by increasing bias but not increasing information. This pattern is seen more clearly in Columns 3 and 4, which consider bottom and top tercile committees separately. In Column 3, the correlation between quality and funding is zero or possibly even negative in committees with low value-added. In contrast, in committees that rank in the top tercile of value-added, the effect of relationships on decision-making that comes through information is positive and significant.

It is important to note that this effect is not a mechanical artifact of the way committee value-added is defined; committees are deemed to perform well if increases in applicant quality translate into increases in funding (see Equation (1.7)). This effect is captured by the coefficient on application quality alone, which indeed is higher for high value-added committees than for low-value-added committees. My results in Table 9 say that, in addition, a high performing committee has more information about a scientist related to a permanent member than one who is not, holding constant their total relatedness to committee members. This is captured by the interaction between application quality and whether an applicant is related to a permanent reviewer. Better performing committees not only have higher correlation between quality and funding overall, but also appear to make more use of the information that permanent members have.

Looking again at Figure 6, ignoring relationships appears to be least harmful in the most poorly performing committees. This is consistent with the finding in Table 9 that bias tends to be higher in poorly performing committees and information tends to be lower. The magnitudes of these effects, however, are not large; regardless of whether relationships are taken into account, the distribution of committee performance is substantial. Understanding other reasons for this dispersion is an important area for future research.

1.8 Conclusion

This chapter develops a conceptual and statistical framework for understanding the tradeoff between bias and information in expert evaluation. In particular, I make use of exogenous variation in reviewer assignments and detailed data on grant application quality to separately identify the effect of bias and information. My results show that, as a result of bias, each additional related permanent reviewer increases an application's chances of being funded by 2.9 percent. Viewed in terms of how committees respond to increases in application quality, being related to a reviewer increases the chances that an application is funded by the same amount as would be predicted by a one-fifth standard deviation increase in its quality. Related reviewers, however, also bring expertise to the committee. I show that their information increases the correlation between quality and funding decisions by over 30 percent. On net, ignoring relationships reduces the quality of the NIH-funded portfolio as measured by numbers of citations and publications by two to three percent.

My results suggest that there may be scope for improving the quality of peer review. I document significant and persistent dispersion in the ability of committees to fund high quality research. Finding ways to eliminate the lower tail of committees, for which increases in quality are actually associated with *decreases* in funding likelihood, could lead to large improvements in the quality of NIH-funded research as measured by citations. The magnitude of these potential benefits are not small when viewed in dollar terms. NIH spending for my sample of approximately 25,000 funded grants totaled over 34 billion dollars (2010 dollars). These grants generated approximately 170,000 publications and 6.8 million citations.[8] This means that, in my sample, the NIH spent about 250,000 dollars per publication or about 5,000 dollars per single citation. Even if these numbers do not represent the social value of NIH-funded research, they suggest that the value generated by high quality peer review can be substantial.

A small part of this overall dispersion can be explained by my finding that high value-added committees extract more information from related reviewers but are less susceptible to bias. Understanding and quantifying other factors affecting committee performance is an important area for future work. Here, the uniformity of NIH's many chartered study sections is helpful because it allows for the possibility of targeted randomized experiments, holding other institutional features constant. For instance, applicants could be assigned to intellectually broad or narrow committees to understand the impact of committee composition on the quality of its decisions. Answers to these questions can provide insights on how to improve project evaluation at the NIH and elsewhere.

[8] I have 170,000 publications linked to grants via formal grant acknowledgments computed from the PubMed database. PubMed, however, undercounts citations because it only counts citations from a subset of articles archived in PubMed Central. To arrive at the 6.8 million citations figure, I use total publications calculated via text-matching (about 100,000 publications) and the total citations accruing to those publications (4.3 million) to compute the average number of citations per publication. I then scale this by the 170,000 publications found in PubMed.

1.9 Appendix A: Proof of Proposition 4.1

Nature has drawn true quality Q^*, and types $Q = \begin{cases} Q_R = Q^* + \varepsilon_R & \text{if } R = 1 \\ Q_{UR} = Q^* + \varepsilon_{UR} & \text{if } R = 0 \end{cases}$

Given this, the Perfect Bayesian equilibrium for this game is characterized by:

1. A set of beliefs that the committee has about true quality Q^* given the message M: $\mu(Q^*|M)$.

2. A message strategy $M(Q)$ for a reviewer, given his or her posterior Q.

3. A decision strategy $D(M)$ for the committee, given the reviewer's message.

These strategies and beliefs must be optimal in the following sense:

1. For each Q^*, $\int_{M \in \mathbf{M}} \mu(Q^*|M)dM = 1$.

2. For each message M, the committee's decision $D(M)$ must maximize its expected payoffs given their beliefs $\mu(Q^*|M)$:

$$D \in \text{argmax} \int_{Q^* \in \mathbf{Q}^*} P^C(D, Q^*)\mu(Q^*|M)dQ^*$$

3. For each posterior Q, the reviewer's message $M(Q)$ must maximize his/her payoffs given the committee's strategy:

$$M \in \text{argmax} \int_{Q^* \in \mathbf{Q}^*} P(D(M), Q^*)f(Q^*|Q)dQ^*, \qquad \text{for } P = \{P^{UR}, P^R\}$$

where $f(\cdot|Q)$ is the density of Q^* given Q.

4. For all reviewer posteriors $Q \in \mathbf{Q}^M$ that induce message M to be sent with positive probability, committee beliefs $\mu(Q^*|M)$ must follow from Bayes' Rule:

$$\mu(Q^*|M) = \frac{\int_{Q^* \in \mathbf{Q}^*} M(Q) f(Q^*|Q) dQ^*}{\int_{Q \in \mathbf{Q}^M} \int_{Q^* \in \mathbf{Q}^*} M(Q) f(Q^*|Q) dQ^* dQ}$$

Having defined the equilibrium concept, I proceed with the proof.

Case 1. Suppose that the reviewer reports her exact posterior and the committee to believes it. In this case, the committee maximizes its utility by funding the proposal if and only if $Q^* + \varepsilon_{UR} > U$. The reviewer has no incentive to deviate from this strategy because she is receiving her highest payoff as well.

Suppose, now, that there were another informative equilibrium. Each message $M \in \mathbf{M}$ induces a probability of funding $D(M)$. Let the messages be ordered such that $D(\mathbf{M}_1) \leq \cdots \leq D(\mathbf{M}_K)$ where $\mathbf{M}_i$ are the set of messages M_i that induce the same probability of funding $D(M_i)$. For reviewers of type $Q^* + \varepsilon_{UR} > U$, the reviewer strictly prefers that the grant be funded. She thus finds it optimal to send the message $\mathbf{M}_K$ that maximizes the probability that the grant is funded. Call this set Y. For $Q^* + \varepsilon_{UR} < U$ the reviewer strictly prefers that the grant be unfunded and sends messages in $\mathbf{M}_1$. Call this set N. The only reviewer who sends any other message is one for which $Q^* + \varepsilon_{UR} = U$. This occurs with probability zero. Thus, with probability one, the space of possible messages is equivalent to $\mathbf{M} = \{Y, N\}$. For this equilibrium to be informative, it must be that $D(N) < D(Y)$.

Given this, the committee's optimal reaction is to fund when $M = Y$ and to reject otherwise. Thus, this equilibrium is payoff equivalent to the first equilibrium. If the we allow uninformative equilibria, $D(\mathbf{M}_1) = \cdots = D(\mathbf{M}_K)$ and any reviewer message is permissible. It must be that $D(M_i) = 0$ for all M_i because the outside option U is assumed to be greater than the committee's prior on quality.

Case 2.

Now consider the case when the reviewer is related and biased. As in Case 1, the set of messages is equivalent, with probability one, to $\mathbf{M} = \{Y, N\}$. In this case, however, reviewers of type $Q^* + \varepsilon_R > U - B$ send $M = Y$ and reviewers of type $Q^* + \varepsilon_R < U - B$ send $M = N$. The only reviewer who sends any other message is one for which $Q^* + \varepsilon_R = U - B$.

Under this strategy, the committee's expectation of Q^* given $M = N$ is $E(Q^*|Q^* + \varepsilon_R < U - B)$. Since this is less than U, the grant goes unfunded. The committee's expectation of Q^* given $M = Y$ is $E(Q^*|Q^* + \varepsilon_R > U - B)$. When this is larger than U, the committee listens to the reviewer's recommendation and we can verify that $D(Y) > D(N)$. There also exists an uninformative equilibria where all grants are rejected.

When $E(Q^*|Q^* + \varepsilon_R < U - B) < U$, the grant is never funded: $D(Y) = D(N) = 0$. In this case, only babbling equilibria exist.

1.10 Appendix B: Proof of Proposition 5.1

Measurement error in Q^* can potentially affect the estimation of α_2 in Equation (1.3). The presence of U, RU, and X, however, will not affect consistency; for simplicity, I rewrite both the regression suggested by the model and the actual estimating equation with these variables partialed out. The remaining variables should then be thought of as conditional on U, RU, and X

$$D = \alpha_0 + \alpha_1 Q^* + \alpha_2 R + \alpha_3 RQ^* + \varepsilon \tag{1.9}$$

$$\begin{aligned} D &= a_0 + a_1 Q + a_2 R + a_3 RQ + e \\ &= a_0 + W + a_2 R + e, W = a_1 Q + a_3 RQ \end{aligned}$$

The coefficient a_2 is given by:

$$a_2 = \frac{\text{Var}(W)\text{Cov}(D,R) - \text{Cov}(W,R)\text{Cov}(D,W)}{\text{Var}(W)\text{Var}(R) - \text{Cov}(W,R)^2} \tag{1.10}$$

Consider $\text{Cov}(W, R)$:

$$\begin{aligned} \text{Cov}(W,R) &= \text{Cov}(a_1(Q^* + v) + a_3 R(Q^* + v), R) \\ &= a_1\text{Cov}(Q^*, R) + a_1\text{Cov}(v, R) + a_3\text{Cov}(RQ^*, R) + a_3\text{Cov}(Rv, R) \end{aligned}$$

Under the assumption that R and Q^* are conditionally independent, this yields:

$$\begin{aligned}
\mathrm{Cov}(W,R) &= a_3\mathrm{Cov}(RQ^*,R) + a_3\mathrm{Cov}(Rv,R) \\
&= a_3\left[E(R^2Q^*) - E(RQ^*)E(R)\right] + a_3\left[E(R^2v) - E(Rv)E(R)\right] \\
&= a_3\left[E(R^2)E(Q^*) - E(R)^2E(Q^*)\right] + a_3\left[E(R^2)E(v) - E(R)^2E(v)\right] \\
&= a_3\left[E(R^2)0 - E(R)^20\right] + a_3\left[E(R^2)0 - E(R)^20\right] \qquad (1.11) \\
&= 0 \qquad (1.12)
\end{aligned}$$

With this simplification, the expression for the estimated coefficient on a_2 becomes:

$$\begin{aligned}
a_2 &= \frac{\mathrm{Var}(W)\mathrm{Cov}(D,R) - \mathrm{Cov}(W,R)\mathrm{Cov}(D,W)}{\mathrm{Var}(W)\mathrm{Var}(R) - \mathrm{Cov}(W,R)^2} \\
&= \frac{\mathrm{Var}(W)\mathrm{Cov}(D,R)}{\mathrm{Var}(W)\mathrm{Var}(R)} \\
&= \frac{\mathrm{Cov}(D,R)}{\mathrm{Var}(R)} \\
&= \frac{\mathrm{Cov}(\alpha_0 + \alpha_1Q^* + \alpha_2R + \alpha_3RQ^* + \varepsilon, R)}{\mathrm{Var}(R)} \\
&= \frac{\alpha_2\mathrm{Var}(R) + \alpha_3\mathrm{Cov}(RQ^*,R)}{\mathrm{Var}(R)} \\
&= \frac{\alpha_2\mathrm{Var}(R) + \alpha_3\left[E(R^2)E(Q^*) - E(R)^2E(Q^*)\right]}{\mathrm{Var}(R)} \\
&= \alpha_2
\end{aligned}$$

1.11 Appendix C: Robustness Checks

Appendix Table A addresses concerns that funding may directly influence the number of citations produced by a grant. Instead of including articles published up to two years after a grant is reviewed, Appendix Table A restricts my analysis to articles published one year before a grant is reviewed up to one year afterward. These publications are highly likely to be based off research that existed before the grant was reviewed. Using this metric, I find nearly identical measures of bias and information.

Another test of my assumption that citations are not directly affected by funding is to ask whether I find bias in the review of inframarginal grants, that is grants that are well above or well below the funding margin. All grants in either group have the same funding status so any bias I find cannot be attributed to differences in funding. Because I hold funding status constant, I can only assess the impact that related permanent members have on an applicant's score not on an applicant's funding status. Appendix Table B reports these results. In Columns 2 and 3, I report estimates of the effect of bias and information in the sample of funded and unfunded grants, respectively. In both cases, I still find evidence that bias exists. One concern is that relationships can still affect funding at the margin. In order to isolate a set of applications for which relationships could not have affected funding status, I consider grants that receive scores well above or well below the payline. Although my estimates on these subsamples are noisier, I still find evidence that bias exists. The magnitudes are somewhat smaller than in my main regression; because these are subsamples, there is no reason to expect that the magnitude of the effect of relationships should be the same for high and low quality grants as it is for the entire sample.

Publications associated with funded grants can also be matched using grant acknowledgments that are recorded in the National Library of Medicine's PubMed database. For the set of funded grants, Appendix Table C reruns my core regressions using citations to publications that explicitly acknowledge a grant as my measure of quality. This analysis differs slightly from my main results using citations because general citations cannot be computed for publications in PubMed. A limited set of citations can, however, be computed using publications in PubMed Central (PMC). PMC contains a subset of life sciences publications made available for free. While this is not as comprehensive a universe as that of Web of Science, it contains, for recent years, all publications supported by NIH dollars. Undercounting of publications would, further, not bias my result as long as it does not vary systematically by whether an applicant is related to a permanent or to a temporary member. I find results that are consistent with my primary findings. In fact, the magnitude of bias I find using explicit grant acknowledgements on the sample of funded grants is the same as the magnitude of bias I find using text-matching publications on this same subsample, as reported in Appendix Table B.

Appendix Table D provides evidence that permanent members do indeed have more influence. In my sample, I observe almost 5,000 reviewers serving both as permanent and as temporary members. For this subset of reviewers, I show that a larger proportion of the applicants whom they have cited are funded when the reviewer is permanent than when the reviewer is temporary, conditional on applicant qualifications. I also show that mean scores for applicants related to a reviewer are higher when that reviewer is permanent. These regressions include reviewer fixed effects.

Appendix Table E adds nonlinearity to Equation (1.6) in order to show that my results are robust to the assumption that error on the reviewer's posteriors in Section 3.3 is uniform. Were ε_{UR} and ε_R distributed otherwise, the association between relatedness and quality would, in general, be nonlinear. To show that this does not make a material difference for my results, I allow for relatedness to permanent reviewers R^P, relatedness to all reviewers R, and quality Q to vary flexibly by including controls for quadratics and cubics in Q, as well as quadratics and cubics of Q interacted with R^P and interacted with R. I find similar results, both qualitatively and quantitatively. In fact, my estimated bias parameter is almost exactly identical.

My results are robust to non-parametric controls for the total number of related applicants (meeting by number of related reviewers fixed effects) and using alternative definitions of relatedness, including using applicant-reviewer mutual citations and citations defined only on publications for which applicants and reviewers are primary authors (first, second, and last position). My results are also robust to alternative identification based on the attendance of reviewers at meetings as opposed to differences between permanent and temporary members. These and other detailed tables are available from the author.

1.12 Appendix D: Estimating Committee Value-Added

I estimate committee value-added using the following regression:

$$D_{icmt} = a + b_{cmt}Q_{icmt} + \mu X_{icmt} + \delta_{cmt} + e_{icmt} \tag{1.13}$$

D_{icmt} is either the actual or counterfactual funding decision for applicant i reviewed during meeting m of committee c in year t. Q_{icmt} is a measure of application quality such as the number of citations it produces in the future and X_{icmt} are detailed controls for the past performance of the applicant, including flexible controls for number of past publications and citations, number and type of prior awarded grants and prior applications, and flexible controls for degrees, gender, and ethnicity. Finally, δ_{cmt} are committee meeting level fixed effects. The coefficients b_{cmt} capture, for each meeting, the correlation between decisions and quality, conditional on X_{icmt}.

Variation in b_{cmt} include sampling error so that $\hat{b}_{cmt}$ is a combination of true value-added plus a noise term. I assume this luck term to be independent and normal:

$$\hat{b}_{cmt} = b^*_{cmt} + \nu_{cmt} \tag{1.14}$$

Under this assumption, $\text{Var}(\hat{b}_{cmt}) = \text{Var}(b^*_{cmt}) + \text{Var}(\nu_{cmt})$ so that the estimate of true variance is upwardly biased from the additional variance arising from estimation error. To correct for this, I note that the best estimate for b^*_{cmt} is given by $\text{E}(b^*_{cmt}|\hat{b}_{cmt}) = \lambda_{ct}\hat{b}_{cmt} + (1-\lambda_{ct})\bar{\hat{b}}_{ct}$ where $\bar{\hat{b}}_{ct}$ is the mean of meeting quality for that committee-year and $\lambda_{ct} = \frac{\sigma^2_{b^*_{cmt}}}{\sigma^2_{b^*_{cmt}} + \sigma^2_{\nu_{cmt}}}$ is a Bayesian shrinkage term constructed as the ratio of the estimated variance of true committee effects, $\sigma^2_{b^*_{cmt}}$, to the sum of estimated true variance $\sigma^2_{b^*_{cmt}}$ and estimated noise variance $\sigma^2_{\nu_{cmt}}$.

To derive this shrinkage term, I use the correlation in meeting quality across the three different funding cycles of a committee fiscal year. In particular, if meeting-specific errors are independent, then $\text{Cov}(\hat{b}_{cmt}, \hat{b}_{cm't}) = \text{Var}(b^*_{cmt}) = \hat{\sigma}^2_{b^*_{cmt}}$. This can be estimated at the committee-year level because a committee meets three times during the year. I construct

$$\hat{\lambda}_{ct} = \frac{\hat{\sigma}^2_{b^*_{cmt}}}{\hat{\sigma}^2_{b^*_{cmt}} + \hat{\sigma}^2_{\nu_{cmt}}} \tag{1.15}$$

so that the adjusted committee value-added is given by:

$$VA_{cmt} = \hat{\lambda}_{ct}\hat{b}_{cmt} \tag{1.16}$$

Because committee membership is not fixed across funding cycles within the same fiscal year (temporary members rotate, permanent members do not), variation in VA_{cmt} represents a conservative lower bound on the variance of committee quality.

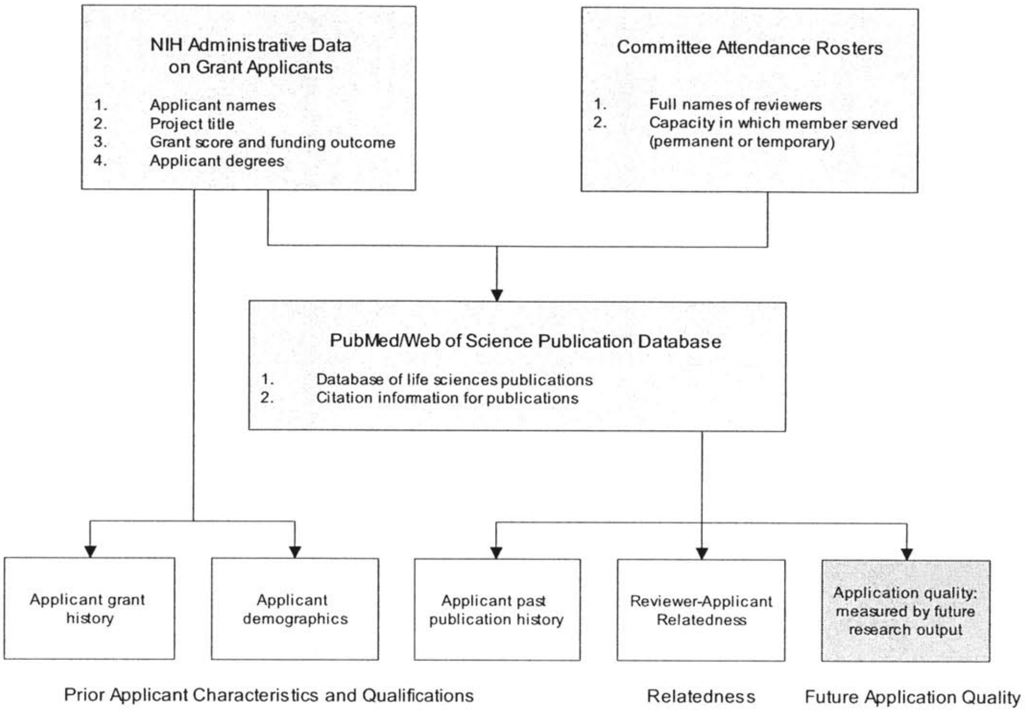
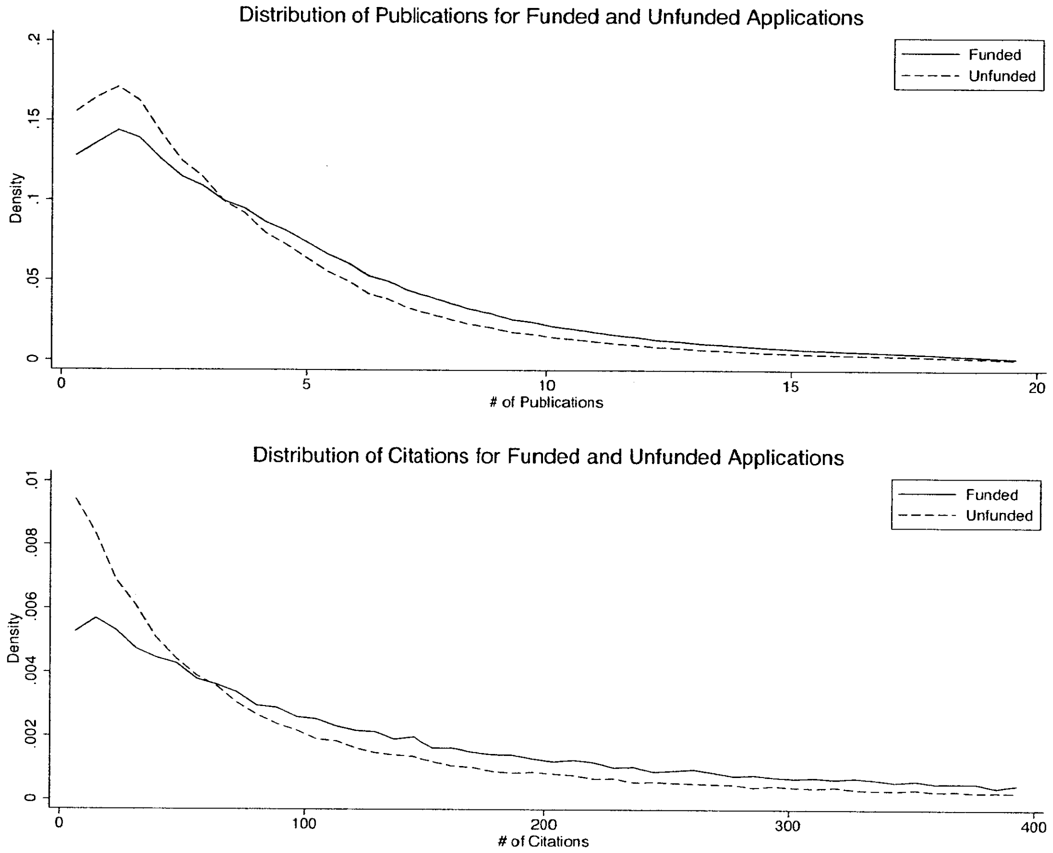

NIH Administrative Data on Grant Applicants
1. Applicant names
2. Project title
3. Grant score and funding outcome
4. Applicant degrees
Committee Attendance Rosters
1. Full names of reviewers
2. Capacity in which member served (permanent or temporary)
PubMed/Web of Science Publication Database
1. Database of life sciences publications
2. Citation information for publications
Applicant grant history
Applicant demographics
Applicant past publication history
Reviewer-Applicant Relatedness
Application quality: measured by future research output
Prior Applicant Characteristics and Qualifications
Relatedness
Future Application Quality

FIGURE 1: DATA SOURCES AND VARIABLE CONSTRUCTION



Distribution of Publications for Funded and Unfunded Applications

Distribution of Citations for Funded and Unfunded Applications

Figure 2: Distribution of application quality: funded and unfunded grants

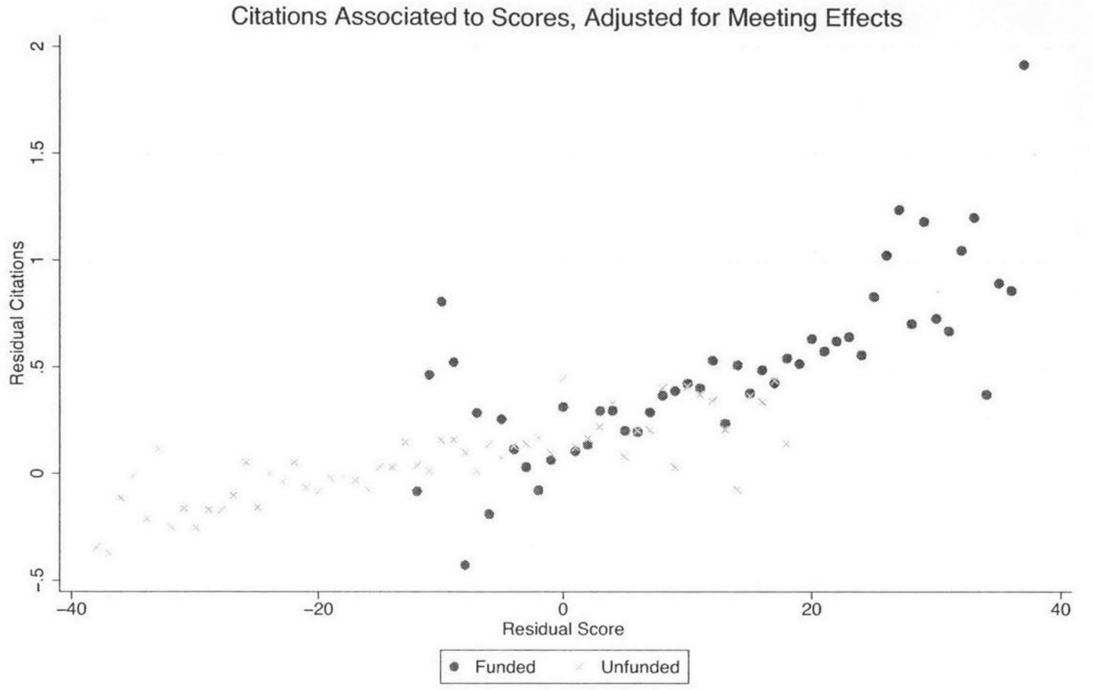


FIGURE 3: MEAN APPLICATION QUALITY BY SCORE: FUNDED AND UNFUNDED GRANTS

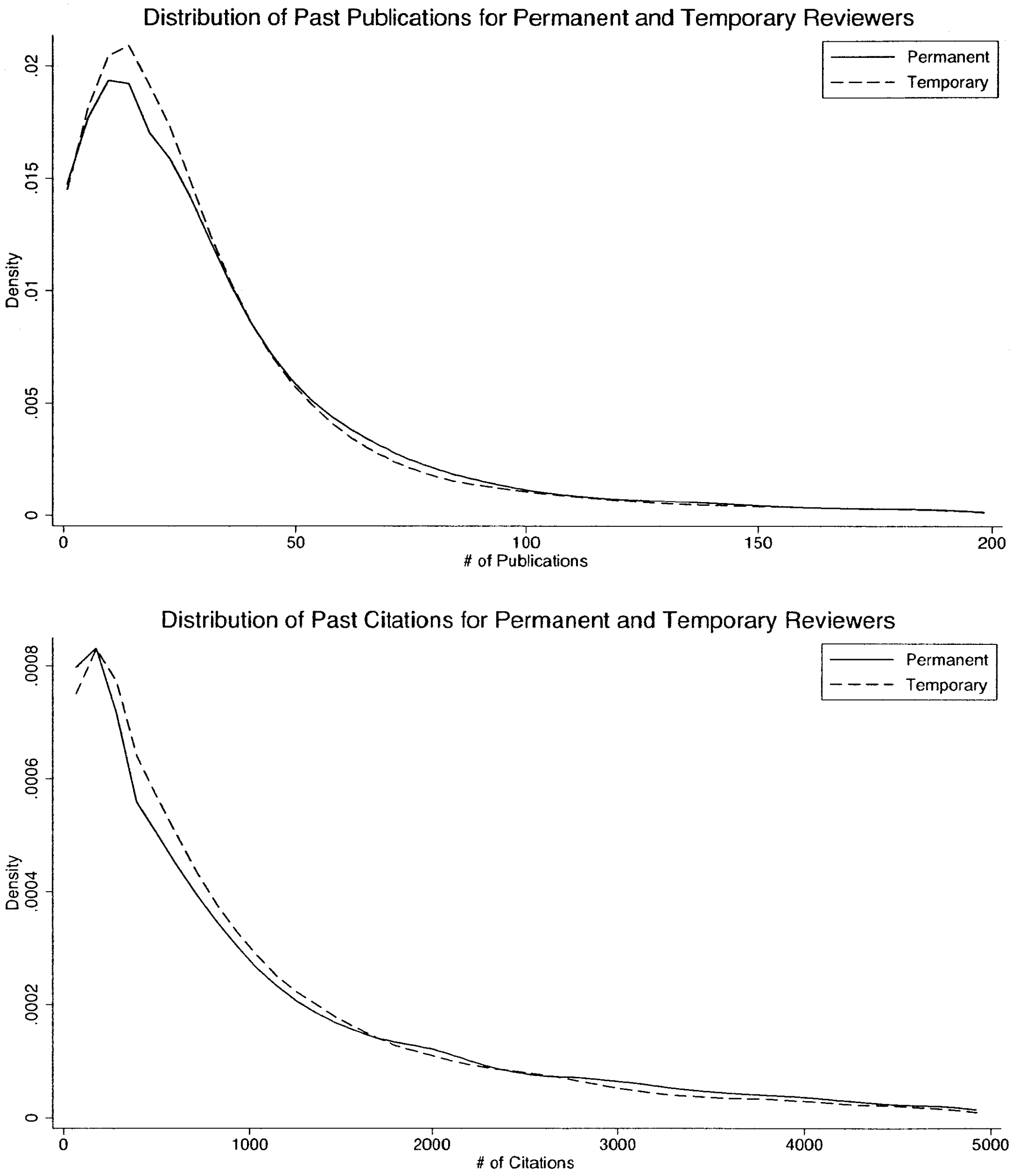


FIGURE 4: DISTRIBUTION OF PAST CITATIONS: PERMANENT AND TEMPORARY REVIEWERS

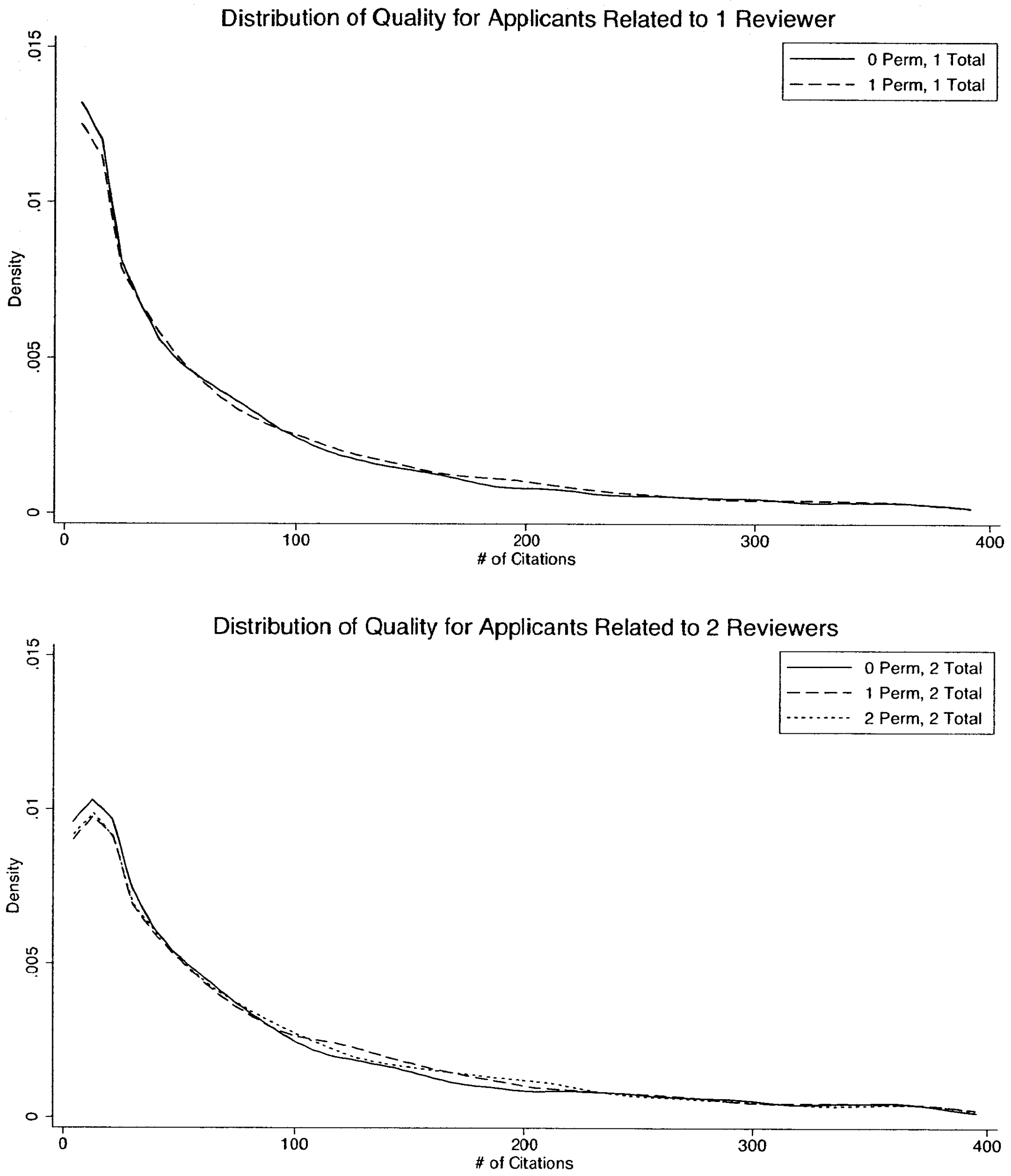


Figure 5: Application quality conditional on total related reviewers

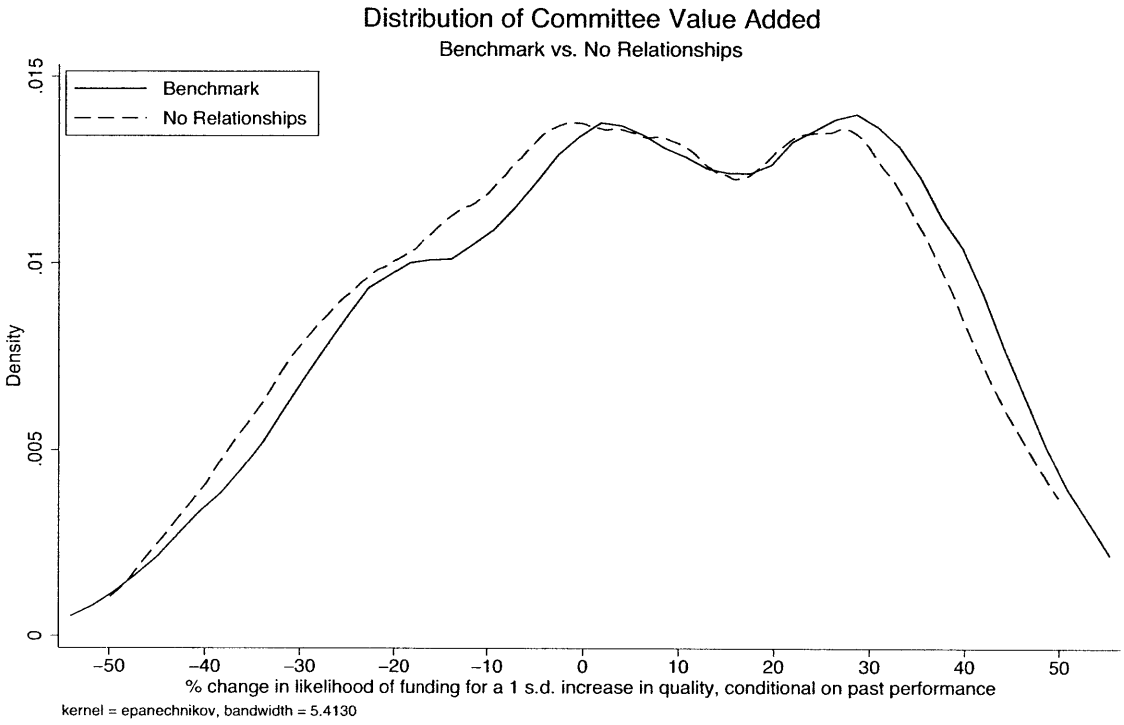


Figure 6: Distribution of meeting-level value-added

Table 1: Applicant Characteristics

	Roster-Matched Sample		Full Sample	
Sample Coverage		Std. Dev.		Std. Dev.
# Grants	93,558		156,686	
# Applicants	36,785		46,546	
Years	1992-2005		1992-2005	
# Study Sections	250		380	
# Study Section Meetings	2,083		4.722	
Grant Characteristics				
% Awarded	26.08		30.48	
% Scored	61.58		64.04	
% New	70.31		71.21	
Percentile Score	70.05	18.42	71.18	18.75
# Publications, grant-publication matched (median)	2	5	2	5
# Citations, grant-publication matched (median)	36	265	38	302
PI Characteristics				
% Female	23.21		22.58	
% Asian	13.96		13.27	
% Hispanic	5.94		5.79	
% M.D.	28.72		29.26	
% Ph.D.	80.46		79.69	
% New investigators	19.70		20.02	
# Publications, past 5 years	15	60	15	55
# Citations, past 5 years	416	1431	423	1474

Notes: The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. Future publications refers to the number of research articles that the grant winner publishes in the 2 years following the grant which share at least one salient word overlap between the grant project title and the publication title. Past publications include any first, second, and last authored articles published in the five years prior to applying for the grant. The full sample includes data from any new or competing R01 grant evaluated in chartered study sections from 1992 to 2005. Investigators with common names are dropped as are any for which the covariates are missing. Social science study sections are dropped.

Table 2: Does being funded directly affect my measure of quality?

Dep var: Grant Quality	(1) No score controls	(2) Controls for smooth function of score
1(Grant is funded)	0.0486*** *(0.0053)*	0.0054 *(0.0104)*
Observations	100276	100276
R-squared	0.3329	0.3335
Past Performance, Past Grants, and Demographics	X	X

Notes: Coefficients are reported from a regession of grant quality on an indicator for whether the grant was funded and controls for applicant characteristics. Column (2) includes controls for quartics in the applicant score. Column (2) compares grant applications with the same score and the same characteristics but which differ in funding status. Applicant characteristics include indicators for sex and whether an applicant's name is Hispanic. East Asian. or South Asian. quartics in an applicant's total number of citations and publications over the past 5 years. indicators for whether an applicant has an M.D. and/or a Ph.D.. and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to.

Table 3: Committee Descriptives

	Roster Matched Sample	
Reviewer Characteristics		Std. Dev.
# Reviewers	18,916	
# Permanent reviewers per meeting	17.23	4.52
# Temporary reviewers per meeting	12.35	7.44
# Meetings per permanent reviewer	3.69	3.03
# Meetings per temporary reviewer	1.78	1.30
# Applications	53.73	17.31
Relationship Characteristics		
# Reviewers who cite applicant	1.94	2.81
# Permanent reviewers who cite applicant	1.11	1.73
# Applicants cited by permanent reviewers	4.12	5.32
# Applicants cited by temporary reviewers	4.12	5.09

Notes: The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005. for which I have study section attendance data. Future publications refers to the number of research articles that the grant winner publishes in the 2 years following the grant which share at least one salient word overlap between the grant project title and the publication title. Past publications include any first. second, and last authored articles published in the five years prior to applying for the grant. Investigators with common names are dropped as are any for which the covariates are missing. Social science study sections are dropped.

TABLE 4: CHARACTERISTICS PERMANENT AND TEMPORARY MEMBERS

	Permanent	**Temporary**
Number of reviewers	9371	14067
Reviewer Characteristics		
% Female	31.68	24.28
% Asian	14.99	13.08
% Hispanic	6.40	5.05
% M.D.	27.42	25.85
% Ph.D.	79.45	80.99
# Publications, past 5 years (median)	22	21
# Citations, past 5 years (median)	606	590

Reviewer Transitions

	% Permanent in the Past	**% Permanent in the Future**	**% Temporary in the Past**	**% Temporary in the Future**
Current Permanent Members	61.87	63.71	38.11	35.45
Current Temporary Members	16.25	41.30	32.73	50.13

Notes: The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. Future publications refers to the number of research articles that the grant winner publishes in the 2 years following the grant which share at least one salient word overlap between the grant project title and the publication title. Past publications include any first, second, and last authored articles published in the five years prior to applying for the grant. Investigators with common names are dropped as are any for which the covariates are missing. Social science study sections are dropped. Transitions are calculated based on whether a reviewer is present in the roster database during the full sample years from 1992-2005. Means are taken for the years 1997 to 2002 in order to allow time to observe members in the past and future within the sample.

Table 5: What is the effect of being related to a reviewer on an applicant's likelihood of funding?

	(1)	(2)	(3)	(4)	(5)	(6)
	1(Score is above the payline)			**Score**		
	Mean = 0.214, SD = 0.410			Mean = 71.18, SD = 18.75		
Related Permanent Reviewers	0.0328***	0.0153***	0.0063***	1.1083***	0.5184***	0.2285**
	(0.0013)	*(0.0012)*	*(0.0020)*	*(0.0542)*	*(0.0517)*	*(0.0926)*
Total Related Reviewers			0.0067***			0.2163***
			(0.0014)			*(0.0601)*
Observations	93558	93558	93558	57613	57613	57613
R-squared	0.0630	0.0947	0.0950	0.1186	0.1433	0.1436
Committee × Year × Cycle FE	X	X	X	X	X	X
Past Performance, Past Grants, and		X	X		X	X

Notes: Coefficients are reported from a regression of committee decisions (score or funding status) on the number of permanent members related to an applicant, controlling for meeting level fixed effects. Column 2 includes indicators for sex and whether an applicant's name is Hispanic, East Asian, or South Asian, quartics in an applicant's total number of citations and publications over the past 5 years, indicators for whether an applicant has an M.D. and/or a Ph.D., and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to. Column 3 includes an additional control for the total number of related reviewers. The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. A reviewer is related to an applicant if the reviewer has cited any of the applicant's previous research in the 5 years prior to grant review.

Table 6: What is the contribution of bias and information?

	(1)	(2)	(3)	(4)
	1(Score is above the payline)		**Score**	
	Mean = 0.214, SD = 0.410		Mean = 71.18, SD = 18.75	
Related Permanent Reviewers	0.0063***	0.0061***	0.2285**	0.2102**
	(0.0020)	*(0.0020)*	*(0.0926)*	*(0.0926)*
1(1+ Related Permanent Reviewers) × Standardized Future Citations		0.0106**		0.2202
		(0.0049)		*(0.2230)*
Standardized Future Citations		0.0315***		1.1674***
		(0.0039)		*(0.1812)*
Total Related Reviewers × Standardized Future Citations		-0.0016**		-0.0524**
		(0.0006)		*(0.0236)*
Total Related Reviewers	0.0067***	0.0072***	0.2163***	0.2403***
	(0.0014)	*(0.0014)*	*(0.0601)*	*(0.0608)*
Observations	93558	93558	57613	57613
R-squared	0.0950	0.0980	0.1436	0.1453
Committee × Year × Cycle FE	X	X	X	X
Past Performance, Past Grants, and Demographics	X	X	X	X

Notes: Coefficients are reported from a regression of committee decisions (score or funding status) on the variables reported, controlling for meeting level fixed effects and detailed applicant characteristics. Column 1 and 3 reproduce Columns 3 and 6 from Table 5. Column 2 and 4 add controls for application quality and application quality interacted with relatedness to permanent and all reviewers. The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. A reviewer is related to an applicant if the reviewer has cited any of the applicant's previous research in the 5 years prior to grant review. Future citations are standardized to be mean zero, standard deviation 1 within each committee-year. Future citations are calculated using all publications by an applicant in the -1 to 2 years after grant review, with text matching. Applicant characteristics include indicators for sex and whether an applicant's name is Hispanic, East Asian, or South Asian, quartics in an applicant's total number of citations and publications over the pa and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to.

TABLE 7: WHAT IS THE CONTRIBUTION OF BIAS AND INFORMATION? HETEROGENEITY IN APPLICANT AND GRANT TYPE

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	1(Score is above the payline)				**Score**			
	Mean = 0.214. SD = 0.410				Mean = 71.18. SD = 18.75			
	New Investigators		Experienced Investigators		New Grants		Renewal Grants	
Related Permanent Reviewers	-0.0023	-0.0023	0.0071***	0.0068***	0.0040*	0.0038	0.0079**	0.0074**
	(0.0056)	*(0.0057)*	*(0.0022)*	*(0.0022)*	*(0.0024)*	*(0.0025)*	*(0.0036)*	*(0.0035)*
1(1+ Related Permanent Reviewers) × Standardized Future Citations		-0.0072		0.0120**		0.0029		0.0189*
		(0.0139)		*(0.0053)*		*(0.0059)*		*(0.0098)*
Standardized Future Citations		0.0361***		0.0305***		0.0311***		0.0234***
		(0.0090)		*(0.0043)*		*(0.0045)*		*(0.0084)*
Total Related Reviewers × Standardized Future Citations		0.0011		-0.0017***		-0.0002		-0.0023**
		(0.0030)		*(0.0007)*		*(0.0008)*		*(0.0009)*
Total Related Reviewers	0.0101***	0.0100***	0.0063***	0.0069***	0.0071***	0.0071***	0.0036	0.0047*
	(0.0036)	*(0.0036)*	*(0.0015)*	*(0.0015)*	*(0.0015)*	*(0.0015)*	*(0.0024)*	*(0.0024)*
Observations	18428	18428	75130	75130	65776	65776	27782	27782
R-squared	0.1768	0.1797	0.0964	0.0992	0.0807	0.0836	0.1622	0.1643
Committee × Year × Cycle FE	X	X	X	X	X	X	X	X
Past Performance. Past Grants. and Demographics	X	X	X	X	X	X	X	X

Notes: See notes to Table 6. Coefficients are reported from a regression of committee decisions (score or funding status) on the variables reported. controlling for meeting level fixed effects and detailed applicant characteristics. Column 1 and 3 reproduce Columns 3 and 6 from Table 5. Column 2 and 4 add controls for application quality and application quality interacted with relatedness to permanent and all reviewers. New Investigators are those who have not received an R01 in the past. New grants are those that are about a new subject. not a renewal of an existing grant.

Table 8: What is the effect of relationships on the quality of research that the NIH supports?

	Benchmark	**No Relationships**
Number of Funded Grants	24,404	24,404
Number of Grants that Change Funding Status	2,166	2,166
Total # Citations	6,680,590	6,547,750
(% change relative to benchmark)		*-1.99*
Total # Publications	149,600	145,331
(% change relative to benchmark)		*-2.85*
Total # in Top 99% of Citations	10,035	9,815
(% change relative to benchmark)		*-2.19*
Total # in Top 90% of Citations	58,149	56,724
(% change relative to benchmark)		*-2.45*
Total # in Top 50% of Citations	132,490	128,980
(% change relative to benchmark)		*-2.65*
Total # Related Applicants Funded	18,059	17,431
(% change relative to benchmark)		*-3.48*

Notes: Benchmark refers to characteristics of grants ordered according to their predicted probability of funding, using the main regression in Table 6 of funding status on relationships and other characteristics. No relationships refers to ordering of grants under the assumption that relatedness to permanent members and relatedness to permanent members interacted with quality do not matter (their coefficients are set to zero). Expected citations are calculated as fitted values from a regression of citations on relationships, past performance, demographics, and meeting fixed effects. The number of projects that are funded is kept constant within meeting. See text for details.

TABLE 9: DO HIGHLY PERFORMING COMMITTEES MAKE BETTER USE OF RELATED REVIEWERS?

	(1)	(2)	(3)	(4)
Dep var: 1(Score > payline) Mean = 0.214, SD = 0.410	**Value-added < Median**	**Value-added > Median**	**Value-added bottom tercile**	**Value-added top tercile**
Related Permanent Reviewers	0.0066	0.0017	0.0044	0.0033
	(0.0043)	*(0.0053)*	*(0.0067)*	*(0.0069)*
1(1+ Related Permanent Reviewers) × Standardized Future Citations	0.0034	0.0123	-0.0081	0.0307**
	(0.0106)	*(0.0101)*	*(0.0114)*	*(0.0142)*
Standardized Future Citations	-0.0073	0.0635***	-0.0126	0.0772***
	(0.0064)	*(0.0087)*	*(0.0076)*	*(0.0143)*
Total Related Reviewers × Standardized Future Citations	0.0007	-0.0039***	0.0009	-0.0064***
	(0.0012)	*(0.0014)*	*(0.0015)*	*(0.0013)*
Total Related Reviewers	0.0051	0.0121***	0.0055	0.0091**
	(0.0032)	*(0.0031)*	*(0.0055)*	*(0.0038)*
Observations	34494	34385	22962	23129
R-squared	0.0842	0.1101	0.0845	0.1173
Committee × Year × Cycle FE	X	X	X	X
Past Performance, Past Grants, and Demographics	X	X	X	X

Notes: Coefficients are reported from a regression of committee decisions (score or funding status) on the variables reported. controlling for meeting level fixed effects and detailed applicant characteristics. The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005. for which I have study section attendance data. I make the additional restriction that the sample be limited to those committees for which I have value-added data. These are typically committees that I observe meeting at least three times. A reviewer is related to an applicant if the reviewer has cited any of the applicant's previous research in the 5 years prior to grant review. Future citations are standardized to be mean zero. standard deviation 1 within each committee-year. Future citations are calculated using all publications by an applicant in the -1 to 2 years after grant review. with text matching. Applicant characteristics include indicators for sex and whether an applicant's name is Hispanic. East Asian. or South Asian. quartics in an applicant's total number of citations and publications over the past 5 years. indicators for whether an applicant has an M.D. and/or a Ph.D.. and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to.

Appendix Table A: What is the contribution of bias and information? Quality measured by publications 1 year before to 1 year after grant

	(1)	(2)	(3)	(4)
	1(Score is above the payline)		**Score**	
	Mean = 0.214, SD = 0.410		Mean = 71.18, SD = 18.75	
Related Permanent Reviewers	0.0063*** *(0.0020)*	0.0062*** *(0.0020)*	0.2285** *(0.0926)*	0.2166** *(0.0925)*
1(1+ Related Permanent Reviewers) × Standardized Future Citations		0.0101** *(0.0048)*		0.1769 *(0.2057)*
Standardized Future Citations		0.0261*** *(0.0038)*		0.9883*** *(0.1687)*
Total Related Reviewers × Standardized Future Citations		-0.0013** *(0.0006)*		-0.0359 *(0.0222)*
Total Related Reviewers	0.0067*** *(0.0014)*	0.0071*** *(0.0014)*	0.2163*** *(0.0601)*	0.2317*** *(0.0609)*
Observations	93558	93553	57613	57608
R-squared	0.0950	0.0976	0.1436	0.1451
Committee × Year × Cycle FE	X	X	X	X
Past Performance, Past Grants, and Demographics	X	X	X	X

Notes: Coefficients are reported from a regression of committee decisions (score or funding status) on the variables reported, controlling for meeting level fixed effects and detailed applicant characteristics. Column 1 and 3 reproduce Columns 3 and 6 from Table 5. Column 2 and 4 add controls for application quality and application quality interacted with relatedness to permanent and all reviewers. The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. A reviewer is related to an applicant if the reviewer has cited any of the applicant's previous research in the 5 years prior to grant review. Future citations are standardized to be mean zero, standard deviation 1 within each committee-year. Future citations are calculated using all publications by an applicant in the -1 to 1 years after grant review, with text matching. Applicant characteristics include indicators for sex and whether an applicant's name is Hispanic, East Asian, or South Asian, quartics in an applicant's total number of citations and publications over the pa and/or a Ph.D., and indicators for the number of past R01and other NIH grants an applicant has won and indicators for how many she has applied to.

Appendix Table B: What is the contribution of bias and information? Inframarginal grant applications

	(1)	(2)	(3)	(4)	(5)
Dep var: Score Mean = 71.18, SD = 18.75	**All**	**Funded**	**Not Funded**	**Well above payline**	**Well below payline**
Related Permanent Reviewers	0.2102**	0.1252*	0.1492*	0.1118*	0.1132
	(0.0926)	*(0.0725)*	*(0.0889)*	*(0.0636)*	*(0.0821)*
Reviewers) × Standardized Future Citations	0.2202	0.3827**	-0.0396	0.0877	0.0642
	(0.2230)	*(0.1748)*	*(0.2410)*	*(0.1658)*	*(0.1939)*
Standardized Future Citations	1.1674***	0.0002	0.4974**	0.1960	0.0746
	(0.1812)	*(0.1382)*	*(0.2031)*	*(0.1323)*	*(0.1561)*
Total Related Reviewers × Standardized Future Citations	-0.0524**	-0.0266	0.0179	-0.0195	0.0029
	(0.0236)	*(0.0178)*	*(0.0261)*	*(0.0162)*	*(0.0216)*
Total Related Reviewers	0.2403***	0.0100	0.1343**	-0.0252	0.0366
	(0.0608)	*(0.0470)*	*(0.0578)*	*(0.0399)*	*(0.0523)*
Observations	57613	24395	33218	14800	22835
R-squared	0.1453	0.1747	0.1880	0.2491	0.7590
Committee × Year × Cycle FE	X	X	X	X	X
Past Performance, Past Grants, and Demographics	X	X	X	X	X

Notes: Coefficients are reported from a regression of committee decisions (score or funding status) on the variables reported, controlling for meeting level fixed effects and detailed applicant characteristics. The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. A reviewer is related to an applicant if the reviewer has cited any of the applicant's previous research in the 5 years prior to grant review. Future citations are standardized to be mean zero, standard deviation 1 within each committee-year. Future citations are calculated using all publications by an applicant in the -1 to 2 years after grant review, with text matching. Applicant characteristics include indicators for sex and whether an applicant's name is Hispanic, East Asian, or South Asian, quartics in an applicant's total number of citations and publications over the past 5 years, indicators for whether an applicant has an M.D. and/or a Ph.D., and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to.

APPENDIX TABLE C: WHAT IS THE CONTRIBUTION OF BIAS AND INFORMATION? EXPLICIT GRANT ACKNOWLEDGEMENTS FOR THE SAMPLE OF FUNDED GRANTS

	(1)	(2)
Dep var: Score Mean = 71.18, SD = 18.75	**Explict Grant Acknowledgements**	
Related Permanent Reviewers	0.1384*	0.1285*
	(0.0724)	*(0.0734)*
1(1+ Related Permanent Reviewers) × Standardized Future Citations		0.0749
		(0.1004)
Standardized Future Citations		0.4806***
		(0.0770)
Total Related Reviewers × Standardized Future Citations		-0.0191*
		(0.0110)
Total Related Reviewers	-0.0074	0.0086
	(0.0456)	*(0.0472)*
Observations	24395	24395
R-squared	0.1743	0.1793
Committee × Year × Cycle FE	X	X
Past Performance, Past Grants, and Demographics	X	X

Notes: Coefficients are reported from a regression of committee decisions (score or funding status) on the variables reported, controlling for meeting level fixed effects and detailed applicant characteristics. The analytic sample includes all awarded R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. A reviewer is related to an applicant if the reviewer has cited any of the applicant's previous research in the 5 years prior to grant review. Future citations are standardized to be mean zero, standard deviation 1 within each committee-year. Future citations are calculated explicit grant acknowlegments. Applicant characteristics include indicators for sex and whether an applicant's name is Hispanic, East Asian, or South Asian, quartics in an applicant's total number of citations and publications over the past 5 years, indicators for whether an applicant has an M.D. and/or a Ph.D., and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to.

Appendix Table D: Do permanent reviewers have more influence?

	(1) **Proportion of Related Applicants who are Funded**	(2) **Average Score of Related Applicants**
Related Reviewer is Permanent	0.003*** *(0.001)*	0.336** *(0.144)*
Observations	15871	15870
R-squared	0.954	0.571
Reviewer FE	X	X
Past Performance, Past Grants, and Demographics	X	X

Notes: This examines how outcomes for related applicants vary by whether the related reviewer is serving in a permanent or temporary capacity. The sample is restricted to 4909 reviewers who are observed both in temporary and permanent positions. An applicant is said to be related by citations if a reviewer has cited that applicant in the 5 years prior to the meeting. Applicant characteristics include indicators for sex and whether an applicant's name is Hispanic. East Asian. or South Asian. quartics in an applicant's total number of citations and publications over the past 5 years. indicators for whether an applicant has an M.D. and/or a Ph.D., and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to.

Appendix Table E: What is the contribution of bias and information? Nonlinear controls for quality and relatedness

	(1)	(2)	(3)	(4)
	1(Score is above the payline)		**Score**	
	Mean = 0.214, SD = 0.410		Mean = 71.18, SD = 18.75	
Related Permanent Reviewers	0.0063***	0.0062***	0.2285**	0.2240**
	(0.0020)	*(0.0021)*	*(0.0926)*	*(0.0948)*
1(1+ Related Permanent Reviewers) × Standardized Future Citations		0.0188**		0.8273**
		(0.0073)		*(0.3900)*
1(1+ Related Permanent Reviewers) × Standardized Future Citations^2		-0.0010		-0.2431
		(0.0047)		*(0.2347)*
1(1+ Related Permanent Reviewers) × Standardized Future Citations^3		-0.0002		0.0141
		(0.0007)		*(0.0300)*
Standardized Future Citations		0.0644***		2.2399***
		(0.0058)		*(0.2997)*
Standardized Future Citations^2		-0.0225***		-0.6377***
		(0.0044)		*(0.2038)*
Standardized Future Citations^3		0.0022***		0.0575**
		(0.0007)		*(0.0281)*
Total Related Reviewers × Standardized Future Citations		-0.0010		-0.0539
		(0.0014)		*(0.0660)*
Total Related Reviewers × Standardized Future Citations^2		0.0006		0.0299
		(0.0006)		*(0.0279)*
Total Related Reviewers × Standardized Future Citations^3		-0.0001		-0.0038
		(0.0001)		*(0.0026)*
Total Related Reviewers	0.0067***	0.0065***	0.2163***	0.2106***
	(0.0014)	*(0.0014)*	*(0.0601)*	*(0.0607)*
Observations	93558	93558	57613	57613
R-squared	0.0950	0.0994	0.1436	0.1464
Committee × Year × Cycle FE	X	X	X	X
Past Performance, Past Grants, and Demographics	X	X	X	X

Notes: The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. A reviewer is related to an applicant if the reviewer has cited any of the applicant's previous research in the 5 years prior to grant review. Future citations are standardized to be mean zero, standard deviation 1 within each committee-year. Future citations are calculated using all publications by an applicant in the -1 to 2 years after grant review, with text matching. Full controls include 1(1+ Related Permanent Reviewers) X Standardized Future Citations in cubics, Total Related Reviewers X Standardized Future Citations in cubics, and Standardized Future Citations in cubics. Applicant characteristics include indicators for sex and whether an applicant's name is Hispanic, East Asian, or South Asian, quartics in an applicant's total number of citations and publications over the past 5 years, indicators for whether an applicant has an M.D. and/or a Ph.D., and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to.

Chapter 2

Gender Bias in NIH Peer Review: Does It Exist and Can We Do Better?

2.1 Introduction

In this chapter, I examine how the US National Institutes of Health (NIH) treats applications from female investigators. In recent years, success rates for NIH grants have fallen to under 18 percent for new applications, raising concerns that scarcity and uncertainty in funding may deter promising graduate students and postdocs from pursuing academic research careers. These concerns may be exacerbated for female scientists who are already less likely to pursue the senior-level academic positions that rely most on external support (CSEPP 2011, NSF 2007, Martinez 2007, and Ceci and Williams 2011). In this context, bias in the peer review process, either perceived or real, is potentially high-stakes and may contribute to the attrition of valuable scientists.

Previous studies of bias in NIH grant review have shown that success rates for men and women are relatively similar while racial and ethnic minorities are less likely to receive grants compared to whites with similar credentials (Ley and Hamilton 2008, RAND 2005). These studies, however, do not provide a conclusive test of bias for two reasons. First, NIH evaluates grant applications on the merit of the specific project that is proposed, not solely on an applicant's past qualifications. Thus, without observing the quality of the grant proposals themselves, one cannot conclude that two applicants with similar publication histories are equally qualified. Second, comparing all R01 grants with one another risks conflating discrimination with other factors. For example, it has been documented that the percentage of women applying for and winning grants has increased over time, even as success rates have been falling for everyone (CSEPP 2011, Fang and Casadevall 2009). Thus, even in a world without bias, women may be less likely to win grants than similarly qualified male scientists simply because women are more represented during periods when funding is scarce. Conversely, current studies may underestimate the extent of bias if, for example, women tend to work in areas of science where success rates are higher.

This study approaches the analysis of discrimination in a new way. I use data on funded grants only. While this has the disadvantage of not allowing me to quantify bias in terms of a female scientist's likelihood of being funded, it allows me to observe, in great detail, the actual quality of the funded proposal. Thus, in addition to using measures of *past* qualifications to account for differences male and female applicants, I can directly control for the quality of a grant application by matching it via grant acknowledgements to the publications and citations it produces in the *future*. Measures of future performance are unlikely to be subject to post-treatment bias because scores are confidential, they do not affect funding, and I restrict my attention to funded grants. Moreover, instead of comparing grants NIH-wide, I compare grants that are evaluated in the same review meeting. To identify bias, I look for systematic differences in the scores assigned to grants whose PIs differ in gender, but 1) which are evaluated by the same people at the same time; 2) whose PIs have similar publication and funding histories; and 3) which eventually produce similarly cited research.

I conduct this analysis using data from 51,353 successful R01 grant applications from 1992 to 2006. The R01 is the NIH's largest investigator-initiated grant program. Study sections assess the merit of applications by assigning them a priority score that is then converted into a percentile ranking. In most cases, proposals are funded in order of their percentile until its designated funding Institute exhausts its budget. The percentile at which this happens is known as the payline. NIH scores work "backward" in the sense that a better score is lower. For ease of exposition, however, I report a grant's percentile to refer to the percentage of applications submitted to that study section which received a worse priority score, so that higher percentiles are better. I measure gender using probabilities constructed from an applicant's full name. While this is not true gender, it is a more accurate measure in the sense that it captures gender as perceived by reviewers who, like me, only have access to names.

My results indicate that women face greater hurdles, especially in the renewal process. For new R01s, gender bias leads women to receive a one-third percentile worse ranking than comparable men; this gap rises to two-thirds for renewal applications. These score-gaps lead to a 1-5 percent decline in the number of women who are funded.

I examine whether NIH can improve gender representation by improving study sections. I collect data on the gender and ethnicity of NIH study section reviewers for about half my sample of grants. For this subsample, I ask whether the composition of the study section influences how applicants are evaluated. I find that the presence of female reviewers attenuates gender bias. Study sections are unbiased when about a third of their members are women. This is evidence that NIH's ongoing efforts to ensure diversity on study sections is having a positive effect on combating bias in grant review.

2.2 Understanding the Gender Gap in Scores

Figures 1 and 2 show that despite substantial gains, women comprised only a third of R01 awardees as of 2007 and, conditional on being funded, received worse percentile rankings. At the same time, however, female R01 awardees also have weaker publication records, as shown in Figure 3. Together these aggregate patterns are inconclusive: women receive fewer grants and worse rankings, but this disparity may reflect underlying differences in the quality of proposed research.

To isolate the effect of bias, I compare percentiles assigned to individual male and female scientists whose grants were reviewed in the same study section meeting. In the raw comparison, women receive percentiles that are on average 0.725 ($P < 0.001$) percentiles worse than men (see Figure 4). Some of this score-gap is attributable to other observable differences between applicants; female investigators tend to be younger and have fewer past high-impact publications. Controlling for past publications, degrees and grant histories reduces the score-gap for women to 0.521 ($P < 0.001$). (See Supporting Materials for a full list of controls).

There are two classes of possible explanations for this remaining disparity. The first is that female investigators receive worse rankings for reasons not justified by the quality of the research they have proposed. The second is that women receive worse rankings because their applications are on average weaker along dimensions that I do not observe. For instance, R01 grants are partially evaluated on the research environment of a proposed project. If women tend to work in smaller universities with fewer resources, then the gender penalty I find may reflect this and not gender per se (see Ceci and Williams, 2011). More generally, study sections do not observe everything about an application's quality and may instead attempt to infer quality based on what they do observe. In this case, because female scientists tend to have fewer qualifications along many observable dimensions, committee members may—potentially correctly—assume that even though two applicants have similar observable qualifications, the research proposed by the female applicant may still be weaker on some unobserved dimension.

These cases can be distinguished from each other by controlling for the future performance of a grant. If male and female grant applicants with identical future grant performance are given systematically different percentile ranks, we can attribute this gap to bias. (See Supporting Information for details).

2.3 Identifying Bias

I test this hypothesis by constructing detailed measures of a grant's future performance. I use data on funded R01s. The performance of a grant once it is funded is a valid measure of the quality of that grant when it is being evaluated because I restrict my sample to funded grants. At the NIH, the score that an application receives only affects its probability of funding and does not affect the amount of funding. Thus, funded grants with better scores should not on average perform better than funded grants with lower scores for any reason other than that they were originally better proposals.

To measure grant quality, I link funded grants with future publications using data on grant acknowledgements from PubMed. Using grant acknowledgements also has the benefit of obviating issues with publication-matching by name, which is particularly problematic for common names. I assess the relative importance of each paper by comparing the number of citations it receives relative to other publications in the same area that are published in the same year. I then construct the following measures of grant quality: total number of future publications, total number of future citations, and the total number of future publications in the 99.9, 99.5, 99, 95, 90, 75, 66, 50, and 25th percentiles of the citation distribution for publications published in the same year. Citations accruing to grants are also computed from PubMed, which only counts citations of papers archived in PubMed Central. In practice, this means that while future publication counts are accurate, citations are undercounted. This measurement error does not lead to bias because it is consistent across applicants evaluated in the same study section. On average, the difference in citation miscounting between two scientists evaluated in the same meeting of the same study section will be zero. Another potential concern with using citations or publications as a measure of quality is that larger grants may potentially support more researchers and thus mechanically generate more publications or citations. I account for this possibility by including controls for the size of amount of funding allocated to the grant. (See Supporting Information for details).

With these controls, I find a gender bias of 0.470 ($P < 0.001$). For competing renewal applications, gender bias is almost twice as large: 0.753 ($P < 0.001$) compared to 0.378 ($P = 0.015$) for new applications. These results indicate that female investigators on average receive a worse percentile even when their research eventually produces similarly cited research. This is evidence that study sections underestimate the quality of female investigators. Score-gaps are graphed in Figure

4.

2.4 How Can Study Section Performance Be Improved?

The results I find may not be representative of all study sections. I obtained attendance rosters for 2,292 study section meetings from 1992-2005 and match grant applications to the study section in which they were reviewed in order to examine whether the demographics of study section members impacts their assessment of applicants. A concern with this type of analysis is that the relationship between female reviewers and bias could be a function of the field. For instance, fields that are friendlier towards women may have more female reviewers and get stronger female applicants. My analysis controls for this possibility by exploiting meeting to meeting changes in the composition of a study section arising from the turnover of members. Specifically, I control for the total number of female reviewers in all three meetings of a study section during a given fiscal year and compare the extent of gender bias in meeting where there are relatively more women.

I find that the presence of women attenuates gender bias. In study sections with no female reviewers, women face a larger 0.683 ($P = 0.021$) percentile bias in applications, but each additional female reviewer decreases this bias by 0.080 ($P = 0.026$) percentiles. Only the presence of women who attend meeting matters; the overall number women women who attend all meetings of a study section (but who may not be present at a particular meeting) does not have an effect on bias. This is consistent with the existence of bias; there is no reason for the quality of proposals from female applicants to be correlated with the number of female reviewers at a particular meeting of a study section.

Gender bias is neutralized when study sections are about one third female, which is nearly the average in my sample of chartered study sections. These results are consistent with recent research demonstrating that female mentors help younger female faculty (Fang and Casadevall 2009) and that female representation is correlated with increased group performance (Woolley et. al. 2010).

2.5 Conclusion

These results show that women on average receive a half percentile worse rank than those of similarly qualified men, leading to a 0.89 to 4.75 percent reduction in the number of female investigators who are funded. This is evidence that study sections make systematic mistakes when judging the quality of female applicants relative to their male peers. I find that problem of gender bias is attenuated by the presence of more female reviewers on study sections. In particular, bias against female applicants is neutralized when a third of study section members are women. This is true for half of study sections in my sample and, moreover, 75 percent of study sections are at least 20 percent female and almost no sections are less than 10 percent female. NIH efforts to promote the representation of women on study sections appears to be an important step toward ensuring that grant review is both fair and perceived as fair.

2.6 Appendix A: Context and Data

My data on R01 grants and their priority scores come from NIH's e-SPA grant database. Each grant observation includes the full name and degree of its primary investigator, the title of the grant project, the Institute to which it was assigned for funding, the study section meeting to which it was assigned for evaluation, the score given by the study section, and the amount of funding the grant received. Attendance rosters were collected for 286 chartered study sections from the NIH Center for Scientific Review. For each study section meeting, I observe the full names of all members who were present. These data serve as the basis for constructing measures of gender, ethnicity, and grant quality.

I match PIs to their prior publications using the Thomson-Reuters Web of Science (WoS) database. From this, I am able to construct the number of publications an applicant in the years prior to submitting their application, their role in those publications (in the life sciences, this is discernible from author position), and the impact of those publications as measured by citations. For instance, I can identify a publication as "high impact" by comparing the number of citations it receives with the number of citations received by other life science articles that were published in the same year. Citations captured in the WoS database include citations from the vast majority of life science publications. Using NIH administrative data, I compute an applicant's past grant history: how many prior new and renewal grants they have received, including non-R01 NIH grants such as post-doctoral fellowships and career training grants. Career age is defined as the time since an investigator received her last degree.

Performance of the actual grant is computed slightly differently. Instead of linking publications to PI names via WoS, I link publications to a specific grant via grant acknowledgement data from the PubMed database. The PubMed grant acknowledgement data allow me to capture the universe of PubMed articles that acknowledge a particular grant but citations accruing to publications that are linked in this way are computed from PubMed Central (PMC), subset of PubMed articles that are available for free. Recent legislation requires that all NIH funded research to be archived in PMC, but this does not apply retroactively, meaning that while the count of future publications associated to a grant is accurate, the count of citations accruing to those publications will be underestimated. This measurement error does not lead to bias because it is consistent across applicants evaluated in the same study section. Thus, on average, the difference in citation miscounting between two scientists evaluated in the same meeting of the same study section will be zero.

Gender is defined probabilistically based on the first name of the PI or reviewer. Investigators and reviewers are assumed to be female if the probability that they are female is greater than one half. Names for which gender probabilities could not be ascertained were dropped (5 percent of sample) and high frequency names were also dropped (10 percent of sample).

2.7 Appendix B: Methods

2.7.1 Identifying Bias

I use regression analysis to assess the extent of gender and ethnicity bias in NIH peer review. The raw score-gap in assigned percentiles is computed from the following regression:

$$R_{ist} = a_0 + a_1 F_{ist} + \delta_{st} + \varepsilon_{ist}. \tag{2.1}$$

Here, the percentile ranking R_{ist} received by applicant i to study section s at time t is modeled as a function of indicator variables for the applicant's gender, F. Fixed effects δ_{st} capture any unobserved differences in how individual study section meetings score grants so that a_1 can be interpreted as the average difference in percentile ranking received by female applicants relative to males who were reviewed in the same meeting of the same study section. Standard errors are clustered at the level of the committee meeting to account for serial correlation in how committees evaluate grants.

To account for differences in qualifications among applicants, I modify Equation (2.3) to include a set X_{ist} of variables describing the applicant's publication history, career age, degrees, and prior grant history, and grant size.

$$R_{ist} = a_0 + a_1 F_{ist} + \mu X_{ist} + \delta_{st} + \varepsilon_{ist}. \tag{2.2}$$

Specifically, X_{ist} includes controls for 1) the total number of citations that the PI received for all publications acknowledging the grant, 2) the total number of publications acknowledging the grant that are in the 99, 95, 90, 80, 70, ..., 10th percentiles of the citation distribution, 3) indicator variables for the number of past successful new and competing R01s and other NIH grants, 4) indicators for career age, 5) indicators for types of degrees, and 6) funding amount.

Given these controls, the coefficient a_1 is interpreted as the percentile difference in scores between female and male applicants who are reviewed by the same study section meeting, who have similar past publications, degrees, and grant histories. Finally, in order to identify the portion of the percentile gap that is attributable to discrimination, I include additional controls Q_{ist} for the future performance of the grant.

$$R_{ist} = a_0 + a_1 F_{ist} + \beta Q_{ist} + \mu X_{ist} + \delta_{st} + \varepsilon_{ist}. \tag{2.3}$$

The set of grant performance measures I use are: 1) the total number of citations that the PI received for all last authored publications published in the five years after receiving the grant and 2) the total number of last authored publications in the 99.9, 99.5, 99, 95, 90, 75, 66, 50, and 25th percentiles of the citation distribution in the five years prior to receiving the grant. Now a_1 measures the role of bias.

2.7.2 How Large Are These Effects?

To assess the consequences of taking gender into account for the number of female investigators who are funded, I construct a counterfactual portfolio of funded grants under the assumption that female investigators are treated the same as male investigators. This rules out both bias as well as different levels of stringency in review. To do this, I generate a hypothetical payline $\bar{R}^0$ such that anywhere from 5 to 25 percent of the grants I observe are cut. Using this new threshold, I calculate benchmark total number of funded women as the number of female PIs for all grants that fall below $\bar{R}^0$ according to their actual percentile rankings. I then generate counterfactual percentiles for each of the cases above based on the estimated coefficients from Equation (2.2):

$$\begin{aligned} R_{ist}^{\text{Benchmark}} &= \widehat{\alpha} + \widehat{\beta}(\text{Applicant is Female}) + \widehat{\mu}X_{ist} + \widehat{\delta}_{st} \\ R_{ist}^{\text{Gender Neutral}} &= \widehat{\alpha} + \widehat{\mu}X_{ist} + \widehat{\delta}_{st} \end{aligned}$$

I rerank grant applications according to its counterfactual score and again consider the number of female investigators for grants falling above the threshold according to both $R_{ist}^{\text{Benchmark}}$ and $R_{ist}^{\text{Gender Neutral}}$.

2.7.3 How Can Study Section Performance Be Improved?

To assess the impact of study section composition on bias, I estimate the following regression model:

$$\begin{aligned} R_{ist} &= a_0 + a_1(\text{Applicant is Female}) + a_2(\#\text{ Female reviewers present at the meeting}) \\ &+ a_3(\text{Applicant is Female}) \times (\#\text{ Female reviewers present at the meeting}) \\ &+ a_4(\#\text{ Female reviewers present at all study section meetings in a fiscal year}) \\ &+ a_5(\text{Applicant is Female}) \times \\ &\quad (\#\text{ Female reviewers present at all study section meetings in a fiscal year}) \\ &+ \gamma Q_{ist} + \mu X_{ist} + \delta_{st} + \varepsilon_{ist}. \end{aligned}$$

This regression holds constant the overall demographics of a study section in a given year and uses variation in the the attendance of female reviewers from meeting to meeting to identify the effect of having additional women in review committees on the extent of gender bias. The coefficients a_4 and a_5 control for how female applicants are generally treated by a particular study section. This could represent the overall female-friendliness of a field. a_1 is the percentile gender gap for study sections with no female reviewers and a_3 identifies the change in the gender gap when the number of women increases as a result of varying attendance in a particular study section.

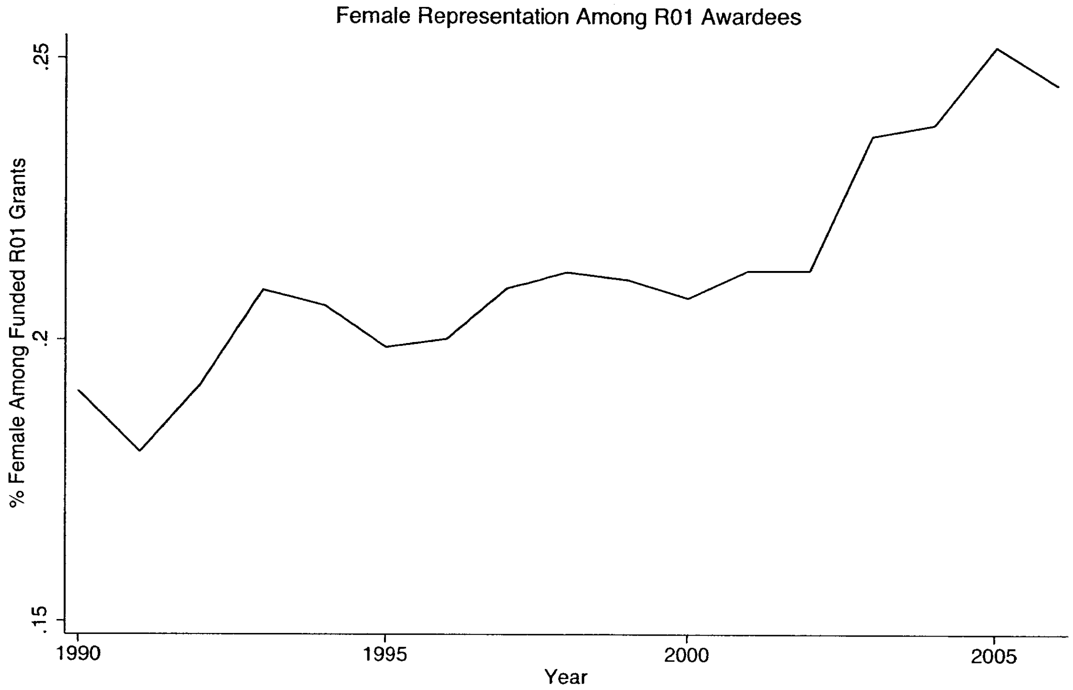


Figure 1: Representation of female investigators among R01 grantees has risen over time, but still remains low.

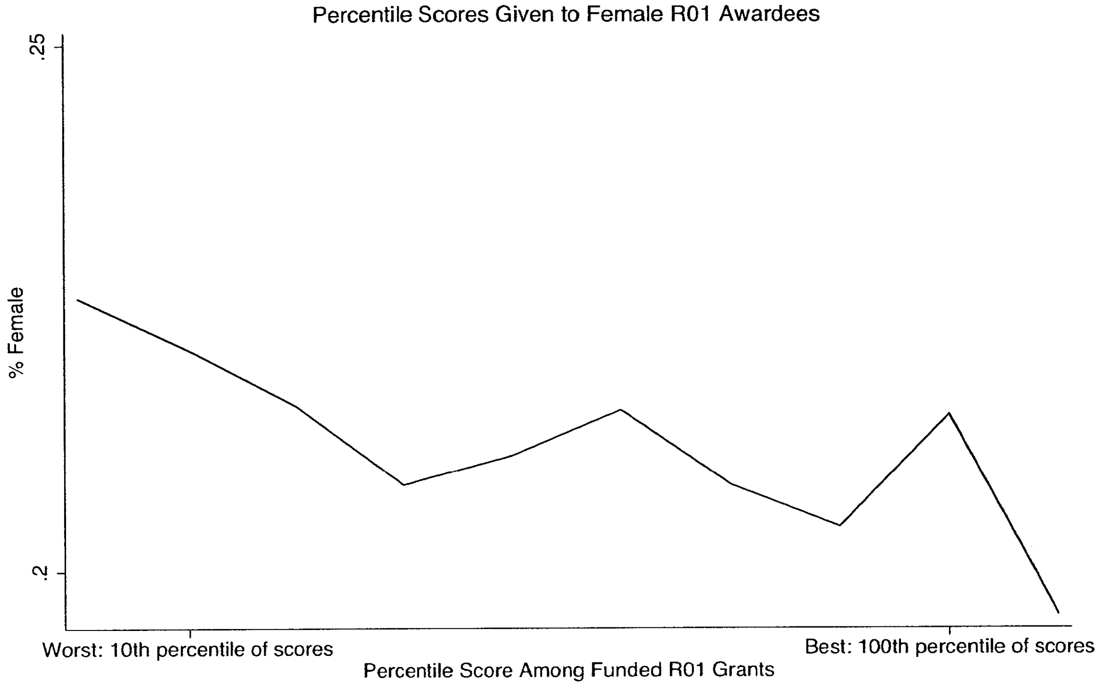


Figure 2: Female investigators are more represented among funded R01 grants with worse percentile rankings.

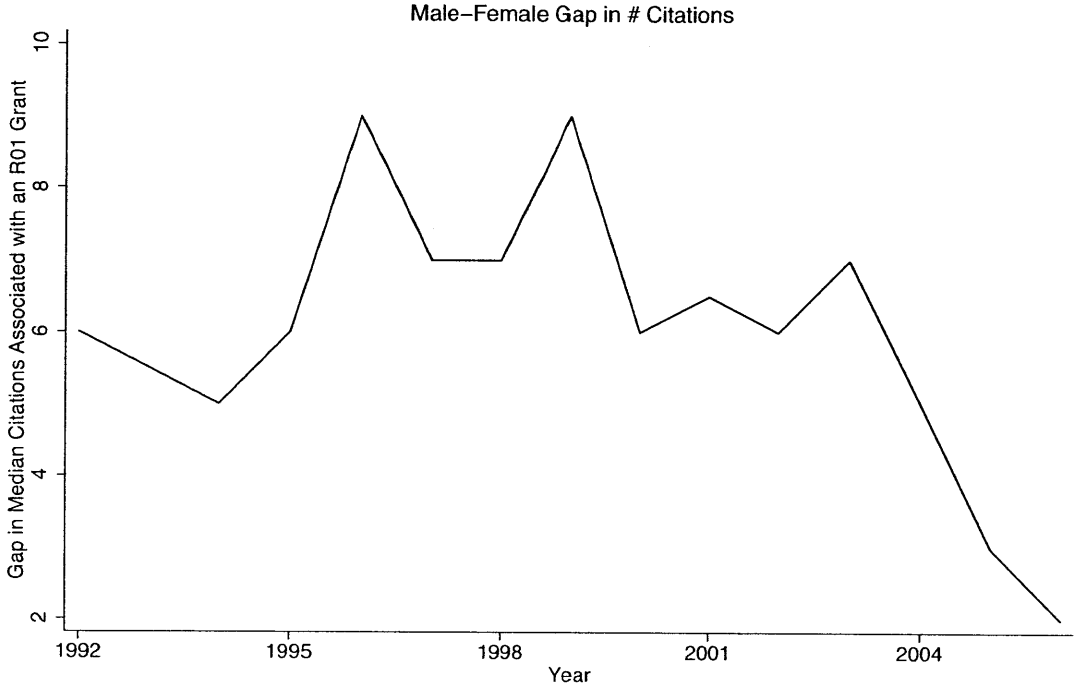


Figure 3: Grants awarded to female investigators are less cited compared with males.

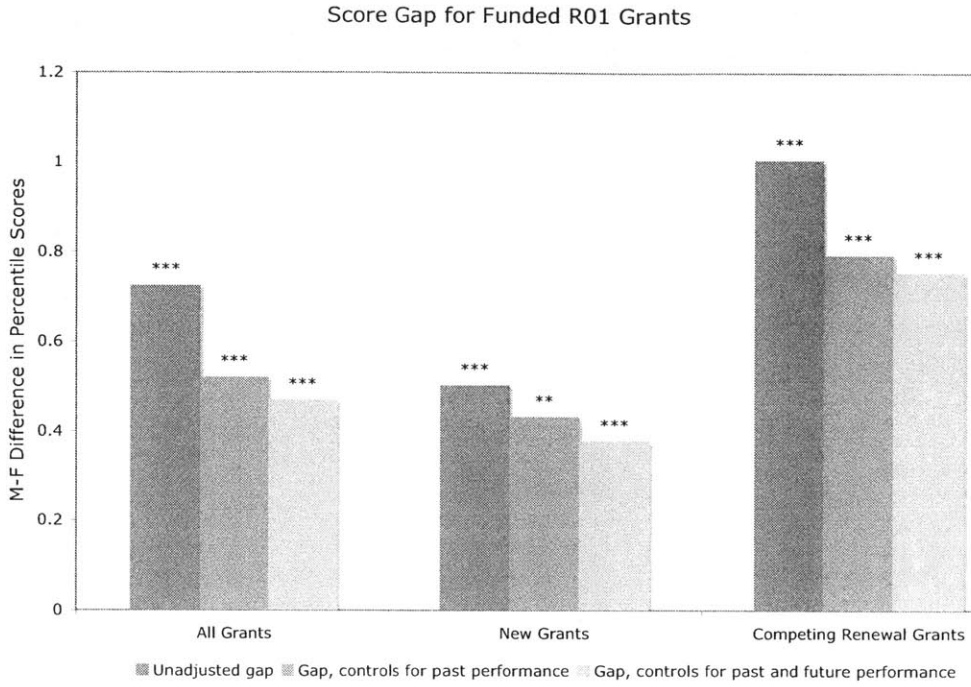


Figure 4: Grants with female PIs receive lower scores than similar grants with male PIs.

TABLE 1: SUMMARY STATISTICS

	Full Sample		Roster Matched Sample	
		SD		SD
Sample Coverage				
# Grants	51,353		25,410	
# Applicants	25,580		16,558	
Years	1992-2006		1992-2005	
# Study Sections	484		285	
# Study Section Meetings	5,480		2,292	
Grant Characteristics				
% New Grants	55.72		54.62	
100-Percentile Priority Score (higher is better)	87.15	9.02	87.13	8.62
# Publications	6.51	7.32	6.83	7.52
# Future Citations (100s)	40.15	85.32	39.57	79.16
PI Characteristics				
% Female	22.21		22.47	
Years since last degree	19.50	9.02	19.81	9.14
% M.D.	27.76		27.54	
% Ph.D.	80.61		81.26	
# Past New or Competing Renewal R01s	4.50	4.37	4.95	4.53
# Total Publications, past 5 years	26.08	45.86	27.23	48.88
# Total Citations, past 5 years	1128	1816	1141	1828
Study Section Characteristics				
# Reviewers			20.233	
% Female			29.78	10.26
# Funded Grants			14.26	5.45

Notes: The full sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2006. The roster matched sample is a subsample that can be matched to the precise meeting (committee and date) in which they were scored. Past publications refers to the number research articles that the grant winner published in the 5 years preceding the grant which fall into the top X-percentile of the citation distribution for research articles published in the same year. Future performance prefers to the performance of publications that acknowledge funding support from the grant.

TABLE 2: UNDERSTANDING GENDER SCORING GAPS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		Full Sample			New Grants			Competing Renewal Grants	
	Controls for Meeting	Controls for Meeting and Past	Controls for Meeting, Past and Future	Controls for Meeting	Controls for Meeting and Past	Controls for Meeting, Past and Future	Controls for Meeting	Controls for Meeting and Past	Controls for Meeting, Past and Future
Female	-0.725*** *(0)*	-0.521*** *(9.23e-07)*	-0.470*** *(9.62e-06)*	-0.503*** *(0.00114)*	-0.432*** *(0.00571)*	-0.378** *(0.0154)*	-1.005*** *(1.08e-08)*	-0.793*** *(8.51e-06)*	-0.753*** *(2.22e-05)*
Observations	51353	51353	51353	28616	28616	28616	22737	22737	22737
R-squared	0.206	0.232	0.237	0.289	0.304	0.308	0.260	0.298	0.304
Meeting FE	X	X	X	X	X	X	X	X	X
Past Performance, Demographics, Degrees, Grant size		X	X		X	X		X	X
Future Performance			X			X			X

Notes: Each column reports results from a separate regression. Sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2006. Past performance includes controls for the number of citations for publications published five years prior to receiving the grant, and the number of publications in the 99, 95, 90, 80, 70,...,10th percentiles of the citation distribution in the five years prior to receiving the grant. Past performance also includes controls for the number of past successful new and competing R01s and other NIH grants. Future performance controls for the number of citations for publications acknowledging the grant, and the number of publications in the 99.9, 99.5, 99, 95, 90, 75, 66, 50, and 25th percentiles of the citation distribution acknowledging the grant. Demographics include controls for career age dummies, ethnicity, gender, and type of degree. Controls are also included for funded grant amount.

TABLE 3: WHAT IF REVIEW WERE GENDER NEUTRAL?

	# Female PIs Funded	
	Benchmark	Gender Neutral
Top 95% Funded		
	10,722	10,817
(% change relative to benchmark)		*0.89*
Top 85% Funded		
	9,358	9,620
(% change relative to benchmark)		*2.80*
Top 75% Funded		
	8,161	8,549
(% change relative to benchmark)		*4.75*

Notes: These are calculated for counterfactual funding thresholds as described in the text. The benchmark is given by fitted scores, not actual scores.

TABLE 4: EFFECT OF STUDY SECTION COMPOSITON ON BIAS: FEMALE INVESTIGATORS

	(1)	(2)
	Baseline	Effect of Female Representation
Female	-0.353** *(0.0134)*	-0.683** *(0.0205)*
Female X (# Study Section Meeting Attendees that are Female)		0.080** *(0.0256)*
Female X (# Study Section Members that are Female--all meetings in a year)		-0.018 *(0.185)*
Observations	25410	25410
R-squared	0.208	0.208
Meeting FE	X	X
Past Performance, Demographics, Degrees, Grant size	X	X
Future Performance	X	X

Notes: Each column reports results from a separate regression. Sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2006. Past performance includes controls for the number of citations for publications published five years prior to receiving the grant, and the number of publications in the 99, 95, 90, 80, 70,...,10th percentiles of the citation distribution in the five years prior to receiving the grant. Past performance also includes controls for the number of past successful new and competing R01s and other NIH grants. Future performance controls for the number of citations for publications acknowledging the grant, and the number of publications in the 99.9, 99.5, 99, 95, 90, 75, 66, 50, and 25th percentiles of the citation distribution acknowledging the grant. Demographics include controls for career age dummies, ethnicity, gender, and type of degree. Controls are also included for funded grant amount.

Chapter 3

Unintended Consequences: No Child Left Behind and the Allocation of School Leaders

3.1 Introduction

The No Child Left Behind Act of 2001 (NCLB) has motivated a vast research program studying the effects of test-based accountability on student performance in U.S. public schools. Though the degree to which test score gains documented at schools under the threat of sanction reflect durable improvements in learning versus strategic behaviors is the subject of debate, one finding of this literature is unambiguous: test-based accountability has significantly changed the incentives and working conditions of teachers and principals.[1] For instance, Reback, Rockoff, and

[1] Drawing on data from both NCLB and smaller state and district-based accountability programs, studies of the effect of accountability on test scores broadly conclude that accountability programs can raise test scores at poorly-performing schools. Figlio and Rouse (2006), West and

Schwartz (2011) find that NCLB accountability pressures lead untenured teachers to work longer hours and feel less secure in their jobs. Yet despite increased scrutiny at disadvantaged schools, principal pay has largely not adjusted to compensate. This relative change in the risk-reward structure of low- versus high-performing schools raises the concern that NCLB might induce effective principals at low-performing schools—who presumably have the option of working elsewhere—to differentially depart these schools. This chapter provides the first quantitative evidence that I am aware of on the important question of how accountability affects the ability of disadvantaged schools to attract and retain high-quality leaders. My results indicate that in evaluating NCLB's impact on students it is important not only to consider short term test score gains but also the long-term allocative effects of increasing accountability without increasing compensation.

The labor market choices of educators is a critical channel by which NCLB may affect school quality in the long run. An influential body of work demonstrates that teacher and principal quality is a major determinant of student learning and that assigning a student to a good educator can matter more for learning than reducing classroom size or increasing classroom resources.[2] Yet unlike the number

Peterson (2006), Rouse et al. (2007), Chiang (2008), Krieg (2008), Neal and Schanzenbach (2007), and Dee and Jacob (2009) all find test score gains of some kind. The nature of these gains is the subject of more debate. Rouse et al. (2007) and Chiang (2008) find persistent gains in math test scores under Florida's state accountability system, but West and Patterson (2006) show that these test score gains do not carry over to NCLB. Krieg (2008), and Neal and Schanzenbach (2007) do find that NCLB increases test scores, but raise concerns that gains are concentrated in the middle of the ability distribution, suggesting that schools ignore low- and high-achieving students in favor of marginal students. Figlio and Getzler (2002), Jacob (2005), and Reback (2006) show that schools remove poorly-performing students from testing pools by reclassifying low-achieving students into special education and Figlio (2006) documents a similar phenomenon where poorly-performing students are subjected to longer disciplinary suspensions near testing dates. Jacob and Levitt (2003) study teacher cheating in pressured schools, and Figlio and Winicki (2005) document calorie inflation.

[2] Aaronson, Barrow, and Sander (2003), Rockoff (2004), and Hanushek, Kain, and Rivkin (2005) study teacher value-added. See Branch, Hanushek, and Rivkin (2009) for a study of principal value-

of computers or teachers per student, the quality of a school's staff cannot simply be assigned. Rather, teachers and principals make choices about where to work and how much effort to exert.

Loeb, Kalogrides, and Horng (2010) and Cullen and Mazzeo (2007) show that principals have preferences over the types of schools they serve are motivated by the opportunity to change schools. In Cullen and Mazzeo's model, career concerns can improve academic performance by creating a competitive environment in which principals exert effort even absent explicit performance bonuses or sanctions. Yet, the effect of increased accountability on a competitive labor market can create unintended consequences for equity. If teachers and principals change jobs in response to the labor market incentives created by NCLB, then NCLB's initial effect on test scores may not reflect its full effects in equilibrium. In particular, labor market sorting may erode the efficacy of NCLB in achieving its stated objectives, which affirm the value of improving access to high-quality schools for disadvantaged children.

By requiring schools to meet the same proficiency targets regardless of prior student performance, NCLB creates wide variation in the likelihood that a school misses performance targets based on factors, such as student demographics, beyond a principal's control.[3] At the same time, principal salaries, which continue to de-

added. Early studies of principals include Eberts and Stone (1988) and Ballou and Podgursky (1995) who study predictors of principal effectiveness. More recently, Knapp et al, Plecki et al., Portin et al., and Copland and Boatright (2006, Wallace Foundation Report) argue that effective principals are able to develop leadership potential among teachers. Jacob and Lefgren (2005) highlight principals' roles in assessing teacher quality, and Jacob (2010) examines the role of principals in firing teachers. Rockoff et. al. (2011) provide more evidence that principals play and important role in evaluating and improving teacher performance.

[3] In the first year of NCLB in North Carolina, where my analysis takes place, the passing thresholds for reading and math were set at, respectively, 68.9 and 74.6 percent proficiency, well above levels typical at schools serving low-income and minority children. Thus, a principal of a school with poorly-performing students would almost surely fail to meet these performance targets, known as Adequate Yearly Progress (AYP), in the first year, and would be subjected to increased administrative burdens as well as to an increased likelihood of facing sanctions in later years.

pend almost entirely on education and experience, have not differentially adjusted to compensate.[4] Though school districts do provide supplements above the standard salary scale, I show that supplements in my sample did not comparatively increase for principals at poorly-performing schools. Thus, NCLB represents a significant, and largely uncompensated change in the risk and amenities associated with working at disadvantaged schools.

Existing studies of the effect of accountability on teacher labor markets have reached mixed conclusions: Clotfelter et al. (2004) find that accountability increases turnover at poorly-performing schools, but Boyd et al. (2008) find that turnover decreases among teachers whose students are subject to testing. This literature does not address, however, whether or in which direction turnover matters for student performance.[5] Increased turnover might signal either that accountability makes it harder to retain effective teachers or that accountability makes it easier to dismiss ineffective ones. Turnover may not affect school quality at all if teachers or principals do not matter for student performance or if it does not change the ultimate composition of school staff.

The key contribution of this chapter is to use outcome-based measures of prin-

Conversely, a principal of a school with high-performing students is likely to pass AYP almost regardless of his or her actions.

[4] A small literature looks at principal pay and incentives. Billger (2007) uses cross-sectional comparisons to show that district sanctions against a school are associated with lower principal pay and mixed results for graduation and retention. Lavy (2008) uses difference-in-differences to estimate the impact of an Israeli program increasing principal pay by 50% and finds significant, though small, gains in student test scores and subjects taken. Besley and Machin (2009) use UK data to show that principal pay and retention responds to performance: pay is linked to publicly observable performance measures and poorly-performing principals face a higher chance of replacement.

[5] A recent paper by Hanushek and Rivkin tie teacher value-added to turnover and find that teachers who leave urban schools tend to be worse than the ones who stay. Their paper focuses primarily on the mobility choices of teachers early in their career as they discover their aptitude for teaching. I instead focus on the mobility decisions of seasoned educators in response to accountability.

cipal quality to examine whether NCLB accountability leads principals to seek less demanding jobs. By tying mobility to quality, I can answer a rich set of questions regarding the impact of NCLB on principals: which types of principals are more likely to switch schools, what kinds of schools do they move to, and what happens to the distribution of principal quality across schools?

To perform this analysis, I estimate principal quality in the period prior to the implementation of NCLB by extracting principal value-added from student test scores. Next, I use variation in school demographics prior to the adoption of NCLB to measure the likelihood that a school will be subject to sanctions. I examine the impact of NCLB on the distribution of principal effectiveness, as well as on changes in mobility patterns, under the assumption that NCLB's accountability provisions should be more binding for principals of schools that are more likely to miss performance targets based on their pre-period demographics.

Consistent with the existing literature, I find that principals matter for performance and that their effectiveness varies significantly across schools. I show that after NCLB high-ability principals at schools more likely to face sanctions for missing performance targets, known as Adequate Yearly Progress (AYP), disproportionately move to schools less likely to face sanctions. These changes in the assignment of schools to principals translate into economically substantive declines in principal effectiveness at schools serving disadvantaged student populations. As a result of NCLB, a standard deviation increase in the likelihood that a school fails AYP leads to a fifth of a standard deviation decrease in average principal effectiveness, as measured by value-added to students' math test scores. These findings are consistent with a model of principal-school matching in which asymmetric changes in the probability that a principal will face performance sanctions, when not fully compensated by changes in pay, lead principals to prefer schools where they are less likely to face sanctions.

A potential limitation of this approach is that principal value-added is estimated only for principals who switch schools between 1995 and 2002, leading to a selected and relatively small sample of principals. Specification checks, as discussed in the empirical appendix, however, indicate that this selection is unlikely to bias estimates of the *differential* impact of NCLB on high- and low-performing schools.

In the next section, I discuss the implementation of NCLB in North Carolina. Section 3 outlines a model of principal-school matching under accountability. Section 4 outlines my econometric methods and estimating equations. Section 5 describes the data and sample construction. Section 6 presents results and Section 7 concludes.

3.2 Institutional Context

No Child Left Behind was signed into federal law in January 2002 with the goal of enabling all children access to a high-quality education as measured by universal proficiency in math and reading by 2014. It mandated annual testing in these subjects for all students in grades 3 through 8, and at least once during high school, starting in the 2002-2003 school year. Under NCLB, schools are designated passing or failing depending on whether they make a performance target known as Adequate Yearly Progress (AYP). Schools are divided into 9 demographic subgroups and AYP requires that students in each subgroup with over 40 members reach a particular threshold for reading and math scores.[6] If only one subgroup fails to make this target, the entire school is declared failing. Starting in 2003-04, schools could also make AYP by showing at least a 10% improvement in scores for every subgroup that still falls

[6] The subgroups are 1) White; 2) Black; 3) Hispanic; 4) Native American; 5) Asian/Pacific Islander; 7) Multiracial; 7) Free/Reduced Price Lunch Students; 8) Limited English Proficient Students; and 9) Students with Disabilities.

below the performance target.

Sanctions associated with failure to make AYP varies across schools and, in particular, depends on whether 1) a school receives federal Title I funds and, if not, whether the school is located in a district which receives Title I funds. Regardless of funding status, NCLB requires that report cards comparing the performance of all schools be made public. Additionally, after two consecutive years of failing to make AYP in the same subject, all schools are required to develop a School Improvement Plan describing strategies that the school will use to meet future performance targets.

The primary bite of NCLB, however, comes at schools which receive federal Title I funds (approximately 50% of schools). Schools are eligible for Title I funding if they serve a large number or high percentage of poor students. For these schools, NCLB created a schedule of sanctions based on the number of consecutive years a school fails to meet AYP in the same subject. AYP designations are determined in the spring and sanctions apply for the following school year:[7]

- First year: There are no official sanctions for the next year, but parents are notified that their child's school is failing.

- Two consecutive years: The school enters the first year of "Title I Improvement" the following school year. In this phase, schools must enable parents to send their children to a non-failing school in the district, unless the school is in a pilot district offering supplemental educational services as the first year option.

- Three consecutive years: The school enters Year 2 of Title I Improvement at the beginning of the next school year and continues to implement school choice

[7] For more details, see the North Carolina Public Schools' NCLB overview: http://www.ncpublicschools .org/nclb/abcayp/overview/ayp

and supplemental educational services.

- Four consecutive years: The school enters Year 3 of Title I Improvement at the beginning of the next school year. School choice and supplemental educational services continue. In addition, the district can pursue "corrective action," meaning that it can replace the principal and teachers, and restructure the curriculum.

- Five consecutive years: The school enters Year 4 of Title I Improvement at the beginning of the next school year. In addition to the sanctions above, the school must devise a contingency plan for restructuring, where the school can be closed, reopened as a charter, privatized, or be taken over by the state.

- Six consecutive years: The school enters Year 5 of Title I Improvement at the beginning of the next school year. Restructuring plans can be implemented.

Non-Title I schools that are located in districts which receive Title I funds (almost all non-Title I schools in my sample) can be sanctioned at the district level if their district fails to meet AYP as a whole. In these cases, however, the primary accountability falls on the superintendent.

Prior to the enactment of NCLB, there were no federally-mandated standards that governed accountability and testing in US public schools. States (and to a lesser extent, districts) had significant purview in designing their own standards for school performance monitoring and accountability. In practice, however, while most states conducted annual testing, very few had explicit consequences associated with poor performance.

North Carolina, the setting for this analysis, was a notable exception because it already had an accountability program in place prior to the introduction of NCLB.

Though that program, the ABCs of Growth, was an early model for NCLB, its benchmarks and requirements differed substantially. In particular, the ABCs emphasized performance as measured by gains in student performance. The ABCs were first implemented for K-8 students in the 1996-97 school year and, initially, schools were given one of the following four designations: 1) Exemplary, for schools whose average test score gains exceeded expected gains by over 10%; 2) Meets Expectations, for schools whose gains meet expectations but which do not exceed them by 10%; 3) No Recognition, for schools that do not meet growth standards, but whose students are more than 50% proficient; 4) Low Performing, for schools that do not meet growth standards, and whose students are less than 50% proficient.[8] Teachers at schools in the top two categories received small bonuses ($1,500 and $750, respectively).

When considering how the implementation of NCLB may have affected principals, the relevant benchmark is how NCLB changed the perception of sanctions relative to the ABCs, not relative to no accountability at all. Importantly, there were no explicit sanctions associated with poor performance and, in practice, the ABC designations were relatively non-binding: in most years, fewer than 1% of schools were designated Low-Performing. In fact, in the 2001-02 school year, on the eve of NCLB's implementation, only 7 schools, or 0.34% of all schools, failed to meet ABC standards. In contrast, in 2002-03, 53% of schools failed to make AYP in the first year, and in 2004-05, almost 10% of schools were subject to official NCLB sanctions. Moreover, school performance on the ABCs was only weakly correlated with AYP performance; 44% of ABC Exemplary schools failed to make AYP and 73% of schools meeting growth expectations under the ABCs failed to make AYP.

[8] Expected gains were calculated by regressing student level gains on student characteristics using 1994 data and then applying the estimated coefficients to data from future years. For more details, see Ladd and Walsh (2002).

In particular, because ABC designations are based on gains in scores, student demographics are a much stronger predictor of AYP performance; the percentage of white and free lunch students, for instance, explains over 28 percent of variation in AYP status, as opposed to just over 6 percent of variation in ABC status. Because schools with many disadvantaged students are significantly more likely to pass ABCs, the implementation of NCLB particularly affects principals of these schools, relative to the ABCs.

Institutional features of the market for school principals play a significant role in shaping how principals respond to these changes in accountability pressure. Principal salaries are set by a statewide schedule that is primarily a function of principal experience, education, and school size. Principals receive the same state wage regardless of school quality, conditional on size. Further, regardless of ability, principals with the same education and experience also receive the same wage, even though studies have found no relationship between principal education and ability, and little relationship between experience and ability beyond the first two years. School districts may provide additional salary supplements for principals, usually around 10% of total pay, which does vary from district to district, but I do not find evidence that districts systematically compensate principals for the quality of the schools at which they work (Appendix Table A).

In North Carolina, principals work on four-year contracts and are not unionized. This increases accountability pressures in the early years of the Title I Improvement phase because districts do not need to wait until the restructuring phase, when principal replacement is explicitly endorsed under NCLB, to act on information about school performance revealed through testing. Further, in contrast to unionized states, there are no strict seniority preferences in hiring; this means that if principals do respond to accountability pressures, mobility may be more related to performance measures as opposed to measures of tenure or experience.

When a principal vacancy is created, districts post this open position and solicit applications.[9] From the pool of applicants, schools pick finalists who are then invited for onsite interviews with the district, school, and school board. Even though principals are officially employed by their local school district, individual schools make offers to candidates and candidates may receive offers from multiple schools in the same district. Importantly, superintendents cannot explicitly transfer principals to other schools within the district and this limits the extent to which principal moves do not reflect optimization by principals given their choice set.

3.3 Theoretical Framework

NCLB imposes sanctions on schools and their leadership when students fail to achieve a certain level of proficiency on annual tests. Because school demographics strongly predict test scores, NCLB changes the implicit costs of working with students from disadvantaged backgrounds. The following simple model illustrates how principal preferences translate into principal-school allocations and looks at the effect of accountability on these allocations. I ignore the effect of accountability on principal retirement or firing in order to focus on its effect on principal-school matching, which my data is better suited to explore.

Consider M schools and $N > M$ potential principals. Student test scores are a function of student ability $\eta_i \sim N(m_s, 1)$ and principal quality $\mu_p \sim U[-\frac{1}{2}, \frac{1}{2}]$. The distribution of student ability is governed by m_s, where for simplicity I assume that a proportion $m_s = 1$ at advantaged schools and $m_s = 0$ at disadvantaged schools.

[9] Increasingly, districts create standing "talent pools" of teachers and administrators interested in principal positions, but this practice was not used during my sample period.

Let $\gamma > 1/2$ be the proportion of schools that are disadvantaged.[10]

Prior to NCLB, schools care about expected test scores minus wages, which due to the rigidity of state salary schedules in North Carolina, are assumed to be fixed. I provide evidence for this assumption in Appendix Table A.

$$\begin{aligned} V_{ps} &= E[\eta_i + \mu_p] - w \\ &= m_s + \mu_p - w \end{aligned}$$

Principal utility is given by:

$$U_{ps} = w + \theta_p m_s + \xi_{ps} \tag{3.1}$$

where, independent of ability μ_p, principals have preferences $\theta_p \sim N(0,1)$ over the type of school at which they work. θ_p can reflect a variety of preferences. Some principals may prefer working with disadvantaged students out of redistributive preferences. Alternatively, if succeeding at disadvantaged schools sends a stronger signal of quality, then principals with stronger desires to advance in the career ladder may have stronger preferences for low-performing schools. ξ_{ps} is an infinitesimal idiosyncratic preference, which ensures that principals have strict preferences over schools, but which does not affect anything else.

Proposition 3.3.1 *There is a unique, stable allocation of principals to schools. Under this allocation, the highest ability principal is matched with his first choice school, the second highest ability principal is matched with her top choice among the remaining vacancies, and so forth until all vacancies are filled.*[11]

[10] Assuming $\gamma > 1/2$ merely says that advantaged schools are more scarce and is done to reduce the number of cases. The results of the model would still obtain if the opposite were true.

[11] See Appendix A for proof. I have assumed that schools observe μ_p, but my results are the

Principals with μ_p greater than some threshold μ_A will receive offers from both types of schools and have the option of working at either. Half the principals, those with $\theta_p > 0$, will choose advantaged schools and the other half will choose disadvantaged schools. Assuming $\gamma > \frac{1}{2}$ so that advantaged schools are scarce, μ_A is determined when advantaged schools fill their vacancies:

$$\frac{N}{2}\left(\frac{1}{2} - \mu_A\right) = (1-\gamma)M \tag{3.2}$$

This yields

$$\mu_A = \frac{1}{2} - \frac{M}{N}2(1-\gamma).$$

Average quality at advantaged schools is then given by:

$$\begin{aligned} Q_A &= \frac{\frac{1}{2} + \mu_A}{2} \\ &= \frac{1}{2} - \frac{M}{N}(1-\gamma) \end{aligned}$$

Disadvantaged schools fill $(1-\gamma)M$ vacancies with principals who choose to work at disadvantaged schools even though they receive other offers. These are the principals with $\theta_p < 0$ and $\mu_p > \mu_A$. Once these vacancies are filled, there are $\gamma M - (1-\gamma)M = (2\gamma - 1)M$ vacancies remaining. Since advantaged schools have filled all their slots, disadvantaged schools fill these vacancies with principals of quality $\mu \in [\mu_A, \mu_B]$, regardless of their preferences. μ_B solves

$$N(\mu_A - \mu_B) = (2\gamma - 1)M$$

same if instead principals are ranked by $E[\mu_p]$. My measure of principal ability is informative of the effect of NCLB on mobility as long as it is related to school's perceptions of principal ability.

or

$$\mu_B = \frac{1}{2} - \frac{M}{N}.$$

Quality at disadvantaged schools is then a weighted average:

$$\begin{aligned} Q_B &= \frac{1}{\gamma}\left[(1-\gamma)\left(\frac{\frac{1}{2}+\mu_A}{2}\right) + (2\gamma-1)\left(\frac{\mu_A+\mu_B}{2}\right)\right] \\ &= \frac{1}{2} - \frac{M}{N}\left(2-\gamma-\frac{1}{2\gamma}\right) \end{aligned}$$

Disadvantaged schools have lower average quality only because advantaged schools are assumed to be scarce: for $\gamma = \frac{1}{2}$, $Q_A = Q_B$. The initial allocation of principals across $\theta - \mu$ space is illustrated in Figure 1. Advantaged schools are filled entirely by principals with $\mu_p > \mu_A$ and $\theta_p > 0$. Disadvantaged schools are filled by principals with $\mu_p > \mu_A$ and $\theta_p < 0$ as well as by any principal with $\mu_p \in [\mu_B, \mu_A)$, regardless of preferences.

Accountability introduces a sanction that principals and schools pay if average test scores fall below a threshold, which I normalize to zero. Principal quality is assumed to affect the test scores of students. A student of ability η_i exposed to a principal of ability θ_p will post a test score of $\eta_i + \theta_p$. In this case, post-accountability principal utility is given by:

$$\begin{aligned} U_{ps} &= w + \theta_p m_s - c\Pr(\eta_i + \mu_p < 0) + \xi_{ps} \\ &= w + \theta_p m_s - c\Phi\left(-\mu_p - m_s\right) + \xi_{ps} \end{aligned}$$

where c is a sanction that a principal pays and Φ is the normal cdf. Similarly, school utility is given by:

$$V_{ps} \;=\; \mu_p + m_s - \tilde{c}\Phi\left(-\mu_p - m_s\right) - w.$$

Because of the threshold nature of accountability, disadvantaged schools now value principal quality more than advantaged schools even though there are no complementarities between principal and school quality in the production of test scores. Taking NCLB's stated goals of increasing minimal competency seriously, it is efficient to allocate better principals to disadvantaged schools where they make a greater contribution toward achieving proficiency.

Sanctions associated with student performance, however, make disadvantaged schools relatively less attractive for all principals. Prior to accountability, principals with $\theta < 0$ preferred disadvantaged schools, but afterward, this threshold is pushed to $\theta < g(\mu_p) < 0$ where $g(\mu_p) = -c[\Phi(-\mu_p) - \Phi(-\mu_p - 1)]$ is the difference in expected sanctions between advantaged and disadvantaged schools. $g(\mu_p)$ is always negative but it is increasing in principal ability; the better a principal, the less she worries about being exposed to sanctions. Principals with $\theta_p \in (g(\mu_p), 0)$ change their preference from disadvantaged schools to advantaged schools because their concerns about sanctions outweigh their devotion to working at disadvantaged schools.

Now, when advantaged schools make an offer to a principal that disadvantaged schools also want, they will expect $1 - \Phi(g(\mu_p)) > 1/2$ of them to accept the offer. Because accountability increases yield, vacancies at advantaged schools fill up faster so that only principals of quality μ'_A receive offers from both types of schools. μ'_A solves

$$N\left[1 - \Phi(g(\mu))\right]\left(\frac{1}{2} - \mu'_A\right) = (1-\gamma)M$$

yielding

$$\mu'_A = \frac{1}{2} - \frac{M}{N}\frac{1-\gamma}{1-\Phi(g(\mu_p))}.$$

Since $1-\Phi(g(\mu_p)) > 1/2$, we can see that $\mu'_A > \mu_A$. Average quality at advantaged schools is given by:

$$Q'_A = \frac{\frac{1}{2}+\mu'_A}{2} = \frac{1}{2} - \frac{M}{2N}\frac{1-\gamma}{1-\Phi(g(\mu_p))} > Q_A$$

Disadvantaged schools receive a lower yield of $\Phi(g(\mu_p))$ so that when advantaged schools have filled up $(1-\gamma)M$ slots, disadvantaged schools have only filled in $N\Phi(g(\mu_p))\left(\frac{1}{2}-\mu'_A\right) = N\left(\frac{1}{2}-\mu'_A\right) - (1-\gamma)M$. Substituting for u'_A, this leaves

$$\gamma M - \frac{\Phi(g(\mu_p))M(1-\gamma)}{1-\Phi(g(\mu_p))} = \frac{M\left[\gamma - \Phi(g(\mu_p))\right]}{1-\Phi(g(\mu_p))}$$

vacancies remaining. These vacancies are filled by principals with quality in (u'_B, u'_A), regardless of preferences:

$$N\left(\mu'_A - \mu'_B\right) = \frac{M\left[\gamma - \Phi(g(\mu_p))\right]}{1-\Phi(g(\mu_p))}.$$

Solving for u'_B yields $u'_B = \frac{1}{2} - \frac{M}{N} = u_B$. This makes sense because the total number of vacancies has not shifted.

Average quality at disadvantaged schools becomes:

$$\begin{aligned} Q'_B &= \frac{1}{\gamma}\left[\left(\frac{\Phi(g(\mu_p))(1-\gamma)}{1-\Phi(g(\mu_p))}\right)\left(\frac{\frac{1}{2}+\mu'_A}{2}\right) + \left(\frac{\left[\gamma-\Phi(g(\mu_p))\right]}{1-\Phi(g(\mu_p))}\right)\left(\frac{\mu'_A+\mu_B}{2}\right)\right] \quad (3.3) \\ &= \frac{1}{2} - \frac{M}{N}\left(\frac{2-\gamma-\frac{\Phi(g(\mu_p))}{\gamma}}{2\left[1-\Phi(g(\mu_p))\right]}\right) \quad (3.4) \end{aligned}$$

When $\Phi(g(\mu_p)) = 1/2$, e.g. when accountability does not diminish the yield for disadvantaged schools, $Q'_B = Q_B$; for lower yields, $Q'_B < Q_B$.

Figure 2 illustrates the shifting distribution of principals across schools after accountability. Disadvantaged schools retain two types of principals after accountability: those with both strong preferences and high-ability who are not deterred by the threat of sanctions $(\mu_p > \mu'_A, \theta_p < g(\mu_p))$, and those who cannot find jobs elsewhere $(\mu_B < \mu_p < \mu'_A)$. Principals with $\mu_p > \mu'_A, g(\mu_p) < \theta_p < 0$ are the principals that switch as a result of accountability. Equation (3.3) is a weighted average of the quality of these groups and captures the intuition that disadvantaged schools are often staffed by a small number of dedicated, high-quality leaders and many more with few other options.

This model makes the following testable predictions:

1. Average principal quality (or perceived quality) declines at disadvantaged schools following the introduction of NCLB.

2. Average principal quality (or perceived quality) increases at advantaged schools following the introduction of NCLB.

3. These effects are greater at schools for which institutionalized sanctions, c and $\tilde{c}$, are greater.

The model does not make an unambiguous prediction about whether high ability principals are more likely to migrate. On the one hand, only principals with quality above u'_A will have the option of moving from disadvantaged to advantaged schools post-accountability. Intuitively, the highest quality principals at disadvantaged schools may not move because they are not worried about sanctions; in this model, however, a subset of them always do because they do not have strong preferences (θ_p negative, but near zero) that would compel them to stay. On the other, this model also predicts that there will be movement among lower quality principals who used to work at advantaged schools but are now forced out as a result of the influx of higher quality principals formerly at disadvantaged schools. Which effect dominates remains an empirical question.

The conceptual framework presented above differs from actual principal-school matching in several ways. First, I have implicitly assumed that principals can be displaced. In reality, this is unlikely to be the case, so that NCLB's impact on mobility and quality is bounded by the number of vacancies. More generally, differences in queues ex ante at different schools affect the extent to which changes in principal preferences translate into assignment of principals to schools. Schools with long queues are less likely to see a change in the average quality of their principals because marginal changes to the applicant pool are less likely to make a difference in terms of who is hired. Engel and Jacob (2011) show that teachers in the Chicago Public School system are more likely to show interest in schools with lower poverty rates. The quality of principals at Title I schools may be more sensitive to accountability pressures both because sanctions are stronger and, potentially, because queues may be shorter. The predictions of this model are also bound by the number of school's in a principal's choice set. Thus, effects may be also be stronger for large and urban districts where principals have a larger choice set of schools. Results by district characteristics are reported in Table 6.

3.4 Empirical Methods

I test the predictions of the model in Section 3.3 by providing estimates of principal quality μ_p based on principal performance in the period prior to the implementation of NCLB. I then identify schools that are likely to fail AYP based on an index of student demographics from 1995 to 2002. Combining these measures of principal and school performance, I examine the effect of NCLB's threat of sanctions on the distribution of principal quality across schools and on principal mobility. I check if larger sanctions c lead to larger declines in principal quality by examining the effect of NCLB on Title I schools, which are subject to official AYP sanctions compared with non-Title I schools, which are not.

Principal quality is difficult to estimate because it requires separating the effect of a principal on student achievement from unobserved neighborhood or school effects. A principal in one school may have advantages over a principal in another that cannot be captured by controls for school budgets or demographics alone: parental motivation, supportive school boards, and local supplies of teachers are all factors that are difficult to control for, but which may substantially impact student performance.

I quantify principal quality using the following model decomposing student performance into individual, school, and principal components using variation from principal mobility across schools:

$$y_{ispt} = \beta_0 y_{isp't} + \beta_1 X_i + \beta_2 X_{st} + \mu_s + \mu_p + \mu_{t\times g} + \varepsilon_{ispt}. \tag{3.5}$$

Here, y_{ispt} is an outcome for student i at school s in year t under principal p, X_{it} are student demographics, X_{st} are time-varying school characteristics, μ_s are school fixed effects, μ_p are principal fixed effects, $\mu_{t\times g}$ are year-grade fixed effects, and ε_{ispt} is an error term. Typically, teacher value-added regressions include controls for lagged scores, but in the case of principals, doing so ignores the cumulative effects of principals over multiple years. Instead, I include controls $y_{isp't}$ for the most recent test score under previous principals p', if available. The inclusion of school fixed effects controls for persistent differences in student and staff quality. School fixed effects and lagged scores for potentially non-random sorting by principals into schools. Year by grade fixed effects control for time-varying differences in testing regimes.

The variance of the measured fixed effects $\hat{\mu}_p$ in Equation (3.5) overstates true variance in principal quality because it reflects both variation in true principal quality and measurement error. Following the spirit of Kane and Staiger (2008), I adjust these estimates using an Empirical Bayes estimator to shrink high variance observations toward the mean: $VA_p = \lambda_p \hat{\mu}_p$, where λ_p is a principal specific shrinkage factor. Details are described in the Appendix.

The principal fixed effects in Equation (3.5) cannot be identified for principals who stay at a school for the entire duration of the sample period because their contribution cannot be distinguished from a school fixed effect. The remaining principals for whom fixed effects can be identified include principals who are only observed in one school, and who stay for a proper subset of the sample period (newcomers or leavers), and those who are observed at multiple schools (switchers). Fixed effects for non-switchers are confounded with time-school-specific effects that may plausibly be attributed to a host of unobservable factors. As such, I focus on switchers only and attribute principal effectiveness to the portion of student achievement that is correlated across schools that a principal is observed in, but which is not explained by other observables and fixed effects.

Identifying principal effects from movers mitigates concerns about conflating school and principal effects, but introduces new selection issues. Principals are not randomly assigned to schools, and if principals systematically move based on the achievement *gains* of students, then the fixed effects estimated in Equation (3.5) may conflate other reasons for changes in performance with true principal effects. Rothstein (2007) shows, in the context of estimating teacher value-added, that test score gains of students can be predicted by the value-added of their future teachers, indicating that teachers are being assigned to classrooms based on student test score gains. Since scores tend to be mean-reverting, a teacher who is assigned to students with high gains in the previous year is unfairly penalized when score gains likely decrease in the current year.

Rothstein's concerns, however, are less of a problem in the context of studying principals. While principals have substantial knowledge about the test scores and other characteristics of students in their own school and may use this information in assigning teachers to classrooms, they have less information about the test score gains of students at other schools and are thus less likely to use this information in their own mobility decisions.

Another concern about principal quality is that it may evolve over time. If much of a principal's true effectiveness comes from learning, this is not reflected in the fixed effect. Instead of including principal fixed effects in (3.5), I could have included principal covariates such as tenure, experience, and education. Previous research on both teachers and principals, however, indicates that the vast majority of variation in educator quality cannot be explained by observables.[12] Thus, I use principal value-added as an imperfect measure of full variation in principal ability.

More generally, this study is concerned about principal quality insofar as it informs the allocative effects of NCLB. As a result, potential bias in value-added is less problematic for three reasons: first, estimates of changes in the assignment of principals to schools based on value-added reflect changes in the true distribution of quality as long as the bias in principal value-added is systematic across principals; second, value-added may be reflective of perceived principal quality and thus be

[12] See Kane, Rockoff, and Staiger (2007) for teachers and Branch, Hanushek, and Rivkin (2009) for principals.

nonetheless informative about the labor market opportunity of principals; and third, mismeasurement of either perceived or true quality biases me away from finding a systematic relationship between mobility and quality as a result of NCLB. It is worth emphasizing the third point here; if Equation (3.5) produced estimates of principal value added that are not reflective of either principal quality or perceived principal quality, then we do not expect the distribution of this measure across high- and low-performing schools to systematically change after the implementation of NCLB.

Using these estimates of principal quality, I next estimate the impact of NCLB on the allocation of principal quality across schools. To conduct this analysis, I exploit exogenous variation in the likelihood, $P = \Phi(-\mu_p - m_s)$, that a principal faces performance sanctions arising from variation in m_s, the baseline ability of students in school s. Schools with low m_s, for instance those serving disadvantaged student populations, have a higher likelihood of facing sanctions, independent of a principal's ability or actions. I quantify the portion of P that is due to m_s alone by estimating the probability that a school fails AYP based on student demographics only. This characterizes a school's probability of failure for which a principal should not, in theory, be penalized:

$$\Phi(\text{fail}_s) = X_s\beta + \varepsilon_s \tag{3.6}$$

where fail_s is an indicator for whether school s would fail AYP in 2001-2002 under 2002-2003 rules based on the number of demographic subgroups in the school, their performance, and the size of those subgroups. I then use a school's demographics prior to 2002 to predict this measure of performance. The covariates X_s include, for each year from 1995 to 2002, cubics for racial composition, proportion of students eligible for free lunch, percentage of students with a parent with some post-secondary education, and school size, with linear effects that are allowed to be different for K-5 schools and non K-5 schools, and the number of students in each particular subgroup. This specification allows both for the proportion of students who belong to a subgroup to impact a school's probability of failure as well as for the size of these subgroups to matter. This is important because a subgroup's performance does not count for AYP if it there are fewer than 40 members of that group. X_s also includes dummies for whether a school is K-5 or urban.[13] The fitted probability of failing becomes my measure of the inherent likelihood of facing sanctions for principals working at each school. Because I am not predicting actual AYP status, which could be influenced by the implementation of NCLB, I treat $\Phi(\text{fail}_s)$ as a known demographic index that describes principals' perceptions of their likelihood of failure in 2002. Standard errors in the case where $\Phi(\text{fail}_s)$ is thought of as a predicted quantity are reported in the appendix tables.

$\Phi(\text{fail}_s)$ indexes a school's exposure to NCLB sanctions and is fixed across schools over time. In reality, however, probabilities of failure change for a school over time either due to changes in student performance or changes in target thresholds so that $\Phi(\text{fail}_s)$ may not necessarily reflect the likelihood of failure for later years in the post-NCLB period. Constructing my measure of failure probability to reflect real probabilities of failure, however, produces a measure of exposure that is endogenous to principal performance. I choose a static measure of likelihood of failure in order to capture the part of NCLB risk that is outside of a principal's control.

Restricting to principals for whom I have estimated pre-period quality and extending the sample period to follow those principals in the post-NCLB years, I ask whether principal quality at disadvantaged schools changes relative to advantaged

[13] For this calculation, there are 14 targets: math and reading targets for Black, White, Hispanic, Asian, Native American, male, female, and all students.

schools following the implementation of NCLB. The estimating equation is given by:

$$\begin{aligned} VA_{pst} &= \alpha_0 + \alpha_1 \Pr(\text{fail})_s \times \mathbb{I}\{\text{year} > 2002\} + \alpha_3 \Pr(\text{fail})_s \\ &\quad + \alpha_4 \mathbb{I}\{\text{year} > 2002\} + X_{pst} + \delta_d + \delta_t + t \times \delta_d + \varepsilon_{pst} \end{aligned} \tag{3.7}$$

where VA_p is estimated principal quality, Pr(fail) is a school's probability of failing AYP, X_{pst} are principal covariates, and δ_d, δ_t, and $t \times \delta_d$ are, respectively, district and year fixed effects, and district linear time trends. District specific time trends allow principal quality among high- and low-performing schools to be on different trends across districts. The possibility that districts are on separate trends is particularly likely in North Carolina, which includes both rural and urban districts with significant variation in racial composition. In this specification, α_1 identifies the effect of NCLB under the assumption that, within districts, high- and low-performing schools are on stable trends in the absence of NCLB.

Using a complementary specification, I also estimate the effect of NCLB on measures of turnover by substituting mobility variables in the left hand side of (3.7) and examining the impact of NCLB on the characteristics of the next school to which principals are assigned. The estimates of NCLB's effect on aggregate principal mobility can be further refined to investigate heterogeneity in principal mobility by ability. I allow the effect of NCLB to differ for principals above and below median estimated quality and test whether high-ability principals are more likely to move, and, conditional on moving schools, what are the characteristics of their new schools.

3.5 Data

I use administrative records from the North Carolina Public School System. These data have been compiled by the North Carolina Education Research Center into student-school and staff-school matched panels spanning the years 1995 through 2007. These data include a unique staff ID that allows me to track principals as long as they move within the state.

3.5.1 Sample Construction

I estimate Equation (3.5) using student level data in the period prior to NCLB, from 1995 to 2002. Figure 3 outlines my procedure for constructing the final analytic sample. From an initial sample of 4,890 full-time principals in schools which employ at most one principal at a time from 1995 to 2002, I match on student test scores and restrict to student-year observations for which 1) I have data on both math and reading test scores for the current and previous year, 2) schools where there are at least two observed principals in the pre-period, and 3) schools where at least one principal is a switcher in the pre-period.[14] For each of these schools, I retain all observations, including those for years in which the school principal is not a switcher. This yields a subsample of 500 schools and 832 principals. In estimating principal fixed effects, I specify that all school fixed effects must be estimated; this allows me to estimate principal fixed effects for 640 principals, of whom 298 are movers.[15]

To study compositional effects, I follow these principals in the post-NCLB years. This initially expands the number of schools in my sample to 596, but I restrict the

[14] Not all years are represented in this dataset because test scores are available only for a subset of years and grades, so that, strictly, a principal must move from one school-year with test scores to another school-year with test scores before 2002.

[15] In a school with two principals, only fixed effects for one principal can be estimated if school fixed effects are also included. The final principal serves as a reference.

sample to schools with standard grades that remain open for the entire sample period from 1995-2007 to avoid spurious mobility effects coming from school openings and closings. Approximately a third of schools are not observed in all years. The final analytic sample includes observations on 214 principals in 383 schools. Each school is observed for an average of six years over the period 1995 to 2007.

3.5.2 Descriptive Statistics

Estimating principal quality from the subset of principals that switch schools prior to the implementation of NCLB creates a measure of quality that is less likely to be contaminated by unobserved school effects. The cost, however, is that this sample of switcher principals may systematically differ in a way that limits the external validity of my estimates.

Table 1 shows summary characteristics of principals and schools in the analytic sample compared to the universe of principals who are in the school system prior to NCLB. There are significant differences between the two samples. Sample principals are observed in my data for approximately 1.5 years more. By construction, all of them have switched schools at least once in my sample period, compared to 66 percent for the universe of principals. Both sets of principals appear to switch at the same time in their careers, early on in their first principalship while they are still under provisional contracts. The schools represented in my analytic sample are on average more likely to fail AYP. Sample schools are also slightly more urban, have higher minority shares, are more likely to receive Title I funds, and are more likely to be K-5 elementary schools. Principal salary and tenure are both slightly lower. These differences are logical since principals of elementary schools and those working in urban districts may plausibly have more nearby employment options.

Table 2 explores in more detail potential differences between principals who switch schools prior to NCLB and the broader universe of principals. Each panel asks whether sample principals are representative of the universe in terms of how their mobility is correlated with school characteristics. The answer is that they appear to be. While sample principals are more likely to switch schools, they do not differentially prefer to leave certain types of schools. For example, Table 2 Panel 2 shows that sample principals are no more or less likely to switch out of a school on the basis of the proportion of white students than the universe of principals. Other issues of sample selection are discussed in the Appendix.

Table 3 reports correlates of the probability of failure measure defined in Equation (3.6) on a selected set of school characteristics (recall that the actual estimation of Pr(Fail) involves demographic subgroup sizes interacted with school level and other variables). Even among this selection of demographics, the R^2 is quite high, indicating that most of a school's probability of failing AYP can be predicted from student demographics alone. The excluded categories are white students and those not eligible for free lunch, so that the coefficients in Table 3 are of the expected sign: poorly-performing schools have more minority and low income students.

3.6 Results

3.6.1 Principal Quality Estimates

I first estimate principal fixed effects from Equation (3.5) and then adjust for measurement error. There is substantial variation in principal quality. Figure 4 plots the estimated distribution of principal quality in math and reading. The dashed line represents principal quality before applying the shrinkage procedure discussed in

Section 3.4. The shrinkage estimator compresses estimates of principal quality and has the greatest effect at the tails of the principal quality distribution. Nonetheless, principal quality, even adjusted for measurement error, remains highly variable: a one standard deviation increase in principal math quality is predicted to increase the math test score of an average student by a fifth of a standard deviation relative to other North Carolina students in that grade and year. These effects are about twice as large as those estimated for teachers, but come from the fact that I allow for principal effects to accumulate over multiple years by only controlling for prior test scores under a different principal. Principal reading quality is closely correlated with math quality (correlation: 0.727), but variation in reading effects is smaller. The variation in principal math and reading performance I estimate is comparable to principals' effects estimated by Branch, Hanushek, and Rivkin (2009) in Texas, which range from 0.17 to 0.27 of a standard deviation of student test scores per one standard deviation in principal quality.

Principal quality varies systematically with school quality as measured by a school's probability of failing AYP. Figure 5 plots the distribution of estimated principal quality in math for schools above and below the median probability of failure before and after the implementation of NCLB. At low-performing schools (Figure 5, top panel), the lower tail of principal math quality shifts further down after 2003, whereas at high-performing schools, the distribution of principal quality remains comparable, or, if anything, improves slightly after NCLB (Figure 5, bottom panel). The distribution of principal quality in reading follows a similar pattern: shifting up after NCLB at high-performing schools, but shifting slightly down at low-performing schools in the same time (Figure 6).

These differences in the distribution of principal quality translate into economically significant differences in the access that various demographic groups have to a high-quality principal. Each cell in Table 4 reports results from a regression of estimated principal quality on school characteristics, controlling only for district fixed effects. I include district fixed effects because principals typically move within the same district, so that these results are informative about the correlation between principal quality and school demographics among schools in the same district, which are more likely to be in a particular principal's choice set. The results in Table 4 indicate that in the pre-period, principal quality is correlated with student performance, but not correlated with student demographics. Only after the introduction of NCLB do students from disadvantaged backgrounds become significantly less likely to attend a school with a high-quality principal. This result is suggestive of an adverse allocative effect of NCLB: by defining AYP in terms of thresholds that are more difficult to meet at schools with more students from disadvantaged backgrounds, NCLB effectively penalizes principals for the demographics of their students. Table 4 indicates that high-quality principals seem to respond to these incentives by choosing to work at schools with fewer disadvantaged students.

3.6.2 Impact of NCLB on Quality

I next examine the effect of NCLB on the allocation of principal quality across schools in more detail. Using estimated principal quality as outcomes in Equation (3.7), I find that NCLB leads to systematic declines in principal math quality at disadvantaged Title I schools, but not at non-Title I schools, which are not directly subject to AYP sanctions. This result is consistent with the model in Section 3.3, which predicts that quality effects are smaller when the probability of facing sanctions is lower.

The results in Column 2 of Table 5 show that at Title I schools a one standard deviation higher likelihood that a school fails AYP (0.363) leads to a $0.363 \times 0.206 = 0.075$ point decline in math effectiveness attributable to NCLB. Given that the standard deviation of principal math ability is 0.217, this translates into a decline in principal math ability of over a third of a standard deviation. Recalling that a one standard deviation higher-ability is associated with a fifth of a standard deviation increase in test scores, this means that given two Title I schools one standard deviation apart in failure probability, students at the worse school are expected to lose approximately 7% of a standard deviation in math test scores as a result of the allocative effect of NCLB. Declines in test scores predicted by changes in principal quality differ from direct estimates of the effect of NCLB on test scores (which tend to be positive) in that they are based entirely on pre-period test scores. These test scores are less likely to be contaminated by concerns about gaming than scores measured after NCLB.

I find that principal quality in reading at Title I schools does not decreases significantly. When compared to non-Title I schools where quality actually increases, however, Title I schools do face a relative decline in principal quality. This suggests that principals who are good at improving reading test scores may not be switching to higher-performing schools, but rather switching to non-Title I schools where the likelihood of sanctions is lower for any level of student performance.

Table 6 examines heterogeneity of the effect of NCLB across districts. Although estimates are more imprecise, I find that the effects of NCLB are stronger in districts where principals are likely to have more mobility—ones with more schools or those in urban areas where schools are closer.[16] This is consistent with the model in which principal mobility drives changes in the distribution of quality.

[16] These results are not driven solely by Charlotte-Mecklenburg and hold up to its exclusion from the sample.

The regressions in Tables 5 and 6 control for district and year fixed effects, principal age and age squared, and a linear district time trend. I include district fixed effects and district by year time trends to allow for the possibility that principal quality among high- and low-performing schools may be on different trends, depending on the district. Given that North Carolina includes both rural and urban districts with significant variation in racial composition, separate trends by district may be likely.

Next I examine the timing of the decline in principal quality. If the decline in principal quality at Title I schools documented in Table 5 occurs as a result of NCLB, relative principal quality should not depart from its trend until after NCLB is signed into law in 2002 or implemented in 2003. Figure 7 plots the effect of NCLB for each year to show that this is indeed the case. For both principal quality in reading and math at Title I schools, there is a break in pre-period trends around the time that NCLB is implemented, consistent with a causal impact of NCLB on the distribution of principal quality. This is supported by the observation that principal quality in non-Title I schools, which are not subject to direct sanctions, does not seem to react to the implementation of NCLB and stays on the same pre-NCLB trends.

The results in Table 5 and Figure 7 indicate that the implementation of NCLB lead to a decline in principal quality at schools most likely to be affected by NCLB's sanction threats. Furthermore, the estimated decline in principal math effectiveness is economically substantial and does not appear to fade out even four years after the implementation of policy.

3.6.3 Impact of NCLB on Mobility

In this section, I analyze possible mechanisms underlying the decline in average principal quality at disadvantaged schools. Table 7 reports estimates of Equation (3.7) where the outcome of interest is a dummy for whether a principal moves to a different school in the next year. I do not find evidence that NCLB increases the aggregate likelihood that principals of high-risk schools switch jobs, either in my analytic sample or in the universe of principals. The final column of Table 7 examines the impact of NCLB on retirement rates for the universe of principals. I do not report estimates for the sample of principals for whom I have estimates of quality because my sample construction method—requiring that principals switch schools at least once during the pre-period—yields artificially low retirement rates before 2003. Column 3 indicates that although retirements increase dramatically after the introduction of NCLB, there does not seem to be a differential effect at schools more likely to fail AYP.

The allocative impact of NCLB as described in my model, however, comes from heterogeneity in principal mobility patterns by ability. Despite no aggregate effect, Table 8 shows that principal ability is indeed linked to subsequent mobility choices; the threat of sanction appears to affect where principals choose to work next, but not whether they decide to switch schools. Specifically, I report coefficients on $\Pr(\text{fail}) \times \mathbb{I}\{\text{year} > 2002\}$ for principals above and below the median quality in math.

Conditional on switching schools, Columns 2 through 5 of Table 8 indicate that, after NCLB, higher-ability principals at poorly performing schools are more likely to move to schools with lower probabilities of failure, more students at grade level, a larger proportion of white students, and non-Title I schools.[17] Consider two

[17] These results hold when principals are split into terciles or quartiles as well.

effective principals who, prior to NCLB, serve at an advantaged and disadvantaged school, respectively. These results say that, after NCLB is implemented, the principal serving at the disadvantaged school is differentially more likely to move to a higher-performing school than the principal serving at the advantaged school, as compared to before NCLB. Because of the change in the difference-in-difference, this cannot merely be attributed to extant patterns of career progression.

One potential benefit of accountability is that increased scrutiny may increase retirement among low-skill principals. Here, my sample selection criterion prevents me from estimating retirement effects, since the sample requirement that principals appear at least once in the post-period mechanically restricts my sample to principals who do not retire in the pre-period. Thus I cannot estimate baseline retirement rates in the pre-period for the sample of principals with quality estimates.

3.6.4 Is the Market More Responsive to "True" Principal Quality or Luck?

So far, I have been treating principal value-added as a noisy measure of true principal quality. However, the results of my model still obtain as long as principal value-added is related to perceived quality, which need not be strongly related to actual ability. Thus, if the labor market conflates true principal quality with qualities of the school that are beyond a principal's control, then a naive measure of principal quality should be more predictive of changes in mobility than the more complicated value-added measure estimated from Equation (3.5). In the case in which a principal's labor market options are determined by perceived and not true ability, estimates of the change in average measured principal quality resulting from NCLB are not necessarily indicative of changes in true principal quality.

Using a measure of principal quality that excludes school fixed effects, I find that the distribution of naive principal "quality" across schools does not change following NCLB (Table 9, Columns 1-3). Although principals with low-quality under this measure experience higher turnover, they do not appear to move to schools that are observably different (Table 9, Columns 4-7). The lack of predictive power for principal quality based on school quality suggests that principals are not broadly credited with the baseline quality of their school. While this does not rule out the possibility that principals are being rewarded for luck, school districts appear to be more sophisticated in their assessments of principal ability.

3.7 Conclusion

Much research on NCLB has focused on school and teacher efforts to increase student test scores. School staff, however, can respond to the pressures of accountability not only by altering the types of effort it puts toward improving test scores, but also by choosing where to work. A primarily contribution of this chapter is to quantify the importance of this mobility response. To do this, I analyze the allocative effects of accountability on the labor market for school principals by examining the impact of NCLB on principal-school matching. I develop a theoretical framework highlighting the consequences of an uncompensated change in the likelihood that a principal faces sanctions on his or her subsequent mobility and test this model using the implementation of NCLB. I find that NCLB leads to declines in the math and, to a lesser extent, reading quality of principals assigned to schools more likely to face sanctions. As predicted by my model, I find that declines in principal ability at disadvantaged schools are caused by the departure of high-quality principals for schools where these principals have a lower likelihood of facing AYP sanctions. More broadly, this chapter shows that in order to evaluate the success of accountability policies such as NCLB, one needs also to consider its impact on incentives in the broader labor market for educators. As a policy that elevates minimal competency to the forefront of educational goals, NCLB demands that greater resources be allocated to students for whom the presence of a high-quality educator may push them pass the proficiency threshold. Implementing NCLB without fully compensating principals for the increased penalties associated with working with disadvantaged student populations, however, leads to the opposite allocative effect. This chapter suggests that districts or policy makers may want to consider the effects of NCLB on the distribution of talent across schools when setting wages or evaluating accountability practices.

3.8 Appendix A: Proof of Proposition 3.1

I first show that there is a unique, stable allocation of principals to schools where the principal with the highest μ is matched to his top choice school, and the principal with the second highest μ is matched to her top choice among the remaining schools, etc., until all vacancies are filled. Assume that the pairings (p_i, s_i) result, where i is the rank of the principal, and s_i is the school chosen by principal i in the manner just described.

Suppose, however, that there exists a blocking pair (p_i, s_j) for $i \neq j$. Because schools share rankings, in order for school j to prefer i, it must be that $i > j$. However, p_i prefers s_i to any s_j for $j > i$ because school j was in p_i's choice set. Thus, it could not be the case that (p_i, s_j) is a blocking pair. This shows that the proposed allocation is stable.

Suppose further that there exists any other stable allocation. This means that for some i, p_i is not matched with s_i. If p_i is matched with s_j for $j < i$, then school j prefers principal j to i. In order for (p_j, s_j) to not be a blocking pair, it must be that principal j is matched to a school he prefers to s_j, call it s_k. For this to be true, it must be that $k < j$. Then, in order for (p_k, s_k) to not form a blocking pair, principal k must be matched to some s_n, $n < k$. Continuing in this way, we reach a contradiction that school s_1 is matched to some principal other than p_1.

If, on the other hand, p_i is matched with s_j for $i < j$, principal i prefers school i. Thus (p_i, s_i) is a blocking pair unless s_i is matched with p_k, $k < i$. The same contradiction follows.

3.9 Appendix B: Value-added Adjustment

Principal fixed effects $\hat{\mu}_p$ estimated from Equation (5) include estimation error so that, ignoring potential bias, $\hat{\mu}_p$ is a combination of the true effect plus a noise term I assume to be independent and normal:

$$\hat{\mu}_p = \mu_p^* + \nu_p \tag{3.8}$$

In this case, $\text{Var}(\hat{\mu}_p) = \text{Var}(\mu_p^*) + \text{Var}(\nu_p)$ so that the estimate of true variance is upwardly biased from additional variance coming from estimation error.[18] To correct for this, I note that the best estimate for μ_p^* is given by $E(\mu_p^*|\hat{\mu}_p) = \lambda\hat{\mu}_p + (1 - \lambda)\overline{\hat{\mu}_p}$ where $\overline{\hat{\mu}_p} = 0$ by design and $\lambda_p = \frac{\sigma_{\mu^*}^2}{\sigma_{\mu^*}^2 + \sigma_\nu^2}$ is a shrinkage term constructed as the ratio of the estimated variance of true principal effects $\sigma_{\mu^*}^2$ to the sum of estimated true variance $\sigma_{\mu^*}^2$ and estimated noise variance σ_ν^2.

In the teacher effects literature, the common solution to this measurement error

[18] For more discussion about the empirical content of value-added measures see Kane and Staiger (2008) and Rothstein (2009).

problem is to use across-time correlation in estimates of teacher fixed effects to construct λ. Applying this approach to principals, however, requires data on multiple principal moves, which happens rarely in practice: in my sample, only 6% of principals move more than once, compared to 30% who move once. However, estimation of principal fixed effects does offer an advantage over estimation of teacher fixed effects in that principals are responsible for the performance of students in many grades. Thus, instead of looking at time-varying correlation in principal quality in order to estimate the true variation in principal effectiveness, I use cross-sectional variation. Specifically, I estimate Equation (5) separately for each grade to obtain an estimate $\hat{\mu}_{pg}$ of principal p's effectiveness in grade g. If grade-specific errors are independent so that $\hat{\mu}_{pg} = \mu_p^* + \nu_{pg}$, then $Cov(\hat{\mu}_{pg}, \hat{\mu}_{p,g-1}) = Var(\mu_p^*)$ where $\hat{\mu}_{p,g-1}$ is the estimate of principal p's effectiveness on the previous grade $g-1$. Thus, my estimate of the true variance of fixed effects is given by $\hat{\sigma}^2_{\mu^*} = Cov(\hat{\mu}_{pg}, \hat{\mu}_{pg-1})$. Thus, I construct

$$\hat{\lambda}_p = \frac{\hat{\sigma}^2_{\mu^*}}{\hat{\sigma}^2_{\mu^*} + \hat{\sigma}^2_{\nu}} \tag{3.9}$$

so that the adjusted fixed effect is given by:

$$VA_p = \hat{\lambda}_p \hat{\mu}_p \tag{3.10}$$

Estimating the variance of true principal ability across grades instead of across years credits principals for high performance that is common across grades. The downside of this approach is that it attributes school-wide common shocks not captured by the school fixed effect to principal performance. If, however, common shocks do indeed create bias in my shrinkage estimator, this bias should be greater in principal quality measured without school fixed effects at all. I check and find that it is not the case.

3.10 Appendix C: Discretionary District Pay

The model presented in Section 3 assumes that schools do not have the flexibility to offer competitive wages to principals. Although the majority of principal pay is set at the state level, there is still the possibility that school districts may be compensating principals who work at poorly-performing schools after NCLB by altering discretionary salary supplements. Although I do not observe individual supplements, I have district-level data on expenditures for salary supplements and the percentage of principals receiving supplements. If principals are being compensated for increased probabilities of failure at certain schools, then supplements at school districts with more poorly-performing schools should rise relative to higher-performing districts in response to NCLB. This change can happen in two ways: first, average principal supplements can increase at poorly-performing districts; or second, if total district funds for supplements do not change, schools may want to reallocate supplements so that the percent of principals receiving supplements should differentially change. I do not find evidence for either of these district responses. In terms of both supplement size and distribution, districts with more schools likely to fail AYP do not seem to behave differently from low-failure districts. If anything, average supplement sizes tend to decrease at high-failure districts, suggesting that the change in the likelihood that a principal is subject to NCLB sanctions is not being fully compensated by pay changes. These results are reported in Appendix Table A.

3.11 Appendix D: Specification Checks

A key concern in estimating Equation (7) is that VA_{pst} is only observed for principals who are movers in the pre-period, which introduces a potentially non-random missing data problem. To see this more clearly, suppose that principal p with observed value-added is observed at school s in year t, but moves to school s' in year $t+1$. In this case, school s is in my analytic sample in year t, but not in year $t+1$ because I am unable to observe the quality of the new principal at school s (unless the new principal it hires is one for which I have estimated value-added).

A concern for my empirical strategy is that, as a result of my sample construction, α_1 in Equation (7) may be capturing changes in the composition of schools I observe in my sample as opposed to true changes in the assignment of principals to schools arising from NCLB, because schools that retain their principal may be unobservably different from schools that do not. This fact alone, however, is not sufficient to generate bias in α_1.

For clarity, consider two schools, A and B, which are identical on observables and which initially both employ principals for whom I have observed value-added, but suppose that school B's principal leaves. In this case, I continue to observe school A's principal in the next year, but I no longer observe the quality of the next principal at school B. If school B's principal left for reasons related to the unobserved quality of the job at B, then the quality of the next principal at school B, which is unobserved, is likely to be different from the quality of the principal at school A, which is observed. This means that the average quality of principals who are observed in the sample is likely to be different from the true average quality of principals, for both the group of low- and high-risk schools. Yet, α_1 captures the difference-in-difference between quality at high- and low-risk schools, before and after NCLB. Thus, in order for this missing data issue to bias α_1, it must be that the bias in observed average quality 1) differs for high and low-risk schools and 2) changes after NCLB. If only 1) is true, then the bias introduced by the sample selection process is captured by the $\Pr(\text{fail})_s$ term in Equation (7). If only 2) holds, then these differences are captured by the time effects in Equation (7).

Conditions 1) and 2) both hold in one of two scenarios. Under the first, high- and low-risk schools must have different probabilities of leaving my sample and, in addition, there must be a change in the extent to which schools staying in the sample unobservably differ from schools that exit. If the degree of selection on unobservables changes, then the difference between observed average quality and true average quality would change before and after NCLB. If high and low-risk schools are equally likely to exit the sample, however, this bias is the same across Pr(fail)$_s$, so that it is captured by the time effects in Equation (7). Conversely, if the likelihood that high and low-risk schools leave the sample is different, then the bias in true and observed principal quality is likely to differ across these types of schools. If, however, there is no change in the degree of selection on unobservables, this bias does not change after NCLB and thus is captured by the Pr(fail)$_s$ term. A similar logic explains the second scenario leading to bias in α_1, which requires both that schools leaving the sample be unobservably different from those that stay, and that there be a differential change in the likelihood that high- and low-risk schools exit the sample after NCLB.

In the case where α_1 is biased, many sensible stories lead to α_1 being too small and bias me away from finding a negative allocative effect of NCLB on high-risk schools. For instance, suppose that after NCLB schools that lose their principals become more undesirable than observably identical schools that retain their principals. If undesirable schools have a harder time attracting high-quality principals, the average quality of principals in schools that are observed is likely to be higher than the true average quality, and this upward bias is likely to be larger the more missing observations there are. Thus if there is more turnover at high-risk schools (and thus more missing observations), the observed difference-in-difference in quality at low- and high-risk schools underestimates the true difference-in-difference because average quality at high-risk schools is more upwardly biased than average quality at low-risk schools.

In Appendix Table B, I examine the degree to which the probability of exiting my sample changes at low- and high-risk schools following NCLB and find no differential effect. There could still be bias if the selection on unobservables of principals as they leave schools changes after NCLB. I have no direct test for this, but in the remaining columns of Table B I show that there does not appear to be differential changes in the selection of schools out of the sample based on observables.

Another way of addressing this missing data problem is to examine only schools that employ a new principal with observed value-added in the next year in which the school is observed. With quality measurements of both the current and next principal, I can ask whether the next principal employed at high-risk schools is more likely to be lower quality after NCLB. In Appendix Table C, I examine the effect on NCLB on the next principal assigned to a school on a restricted sample where I observe both the quality of the current and future principal. This reduces the sample size by a significant amount, but I find evidence that both the math and reading quality of the next principal falls at high-risk schools following NCLB.

Appendix Table D reports the same results as Table 5 with bootstrapped standard errors in the case when the probability of failure is treated as an estimated quantity and I find similar results. Table E of the appendix reports results using alternative measures of school performance and principal quality. I find that I obtain similar results when measuring a school's exposure to NCLB by using the percentage of AYP targets it is likely to fail, and when I use unshrunken estimates of principal value-added. I find qualitatively similar but smaller and statistically insignificant effects of NCLB when principal value-added is measured using lagged test scores in the previous year. This result may reflect the fact that controlling for lagged test scores for the previous year does not fully credit a principal with cumulative test score gains made in her school.

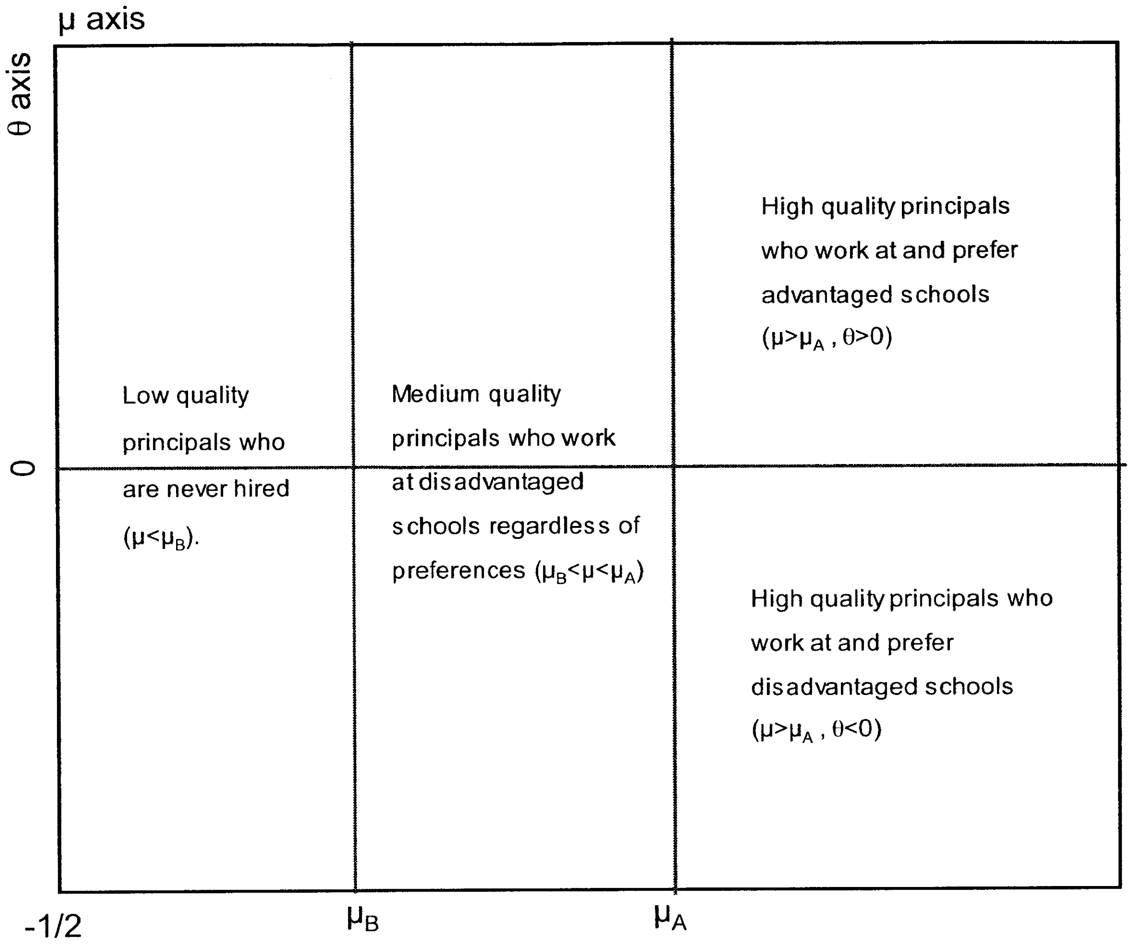

μ axis
θ axis
0
-1/2
μ_B
μ_A
High quality principals who work at and prefer advantaged schools (μ>μ_A , θ>0)
Low quality principals who are never hired (μ<μ_B).
Medium quality principals who work at disadvantaged schools regardless of preferences (μ_B<μ<μ_A)
High quality principals who work at and prefer disadvantaged schools (μ>μ_A , θ<0)

Figure 1: Initial allocation of principals

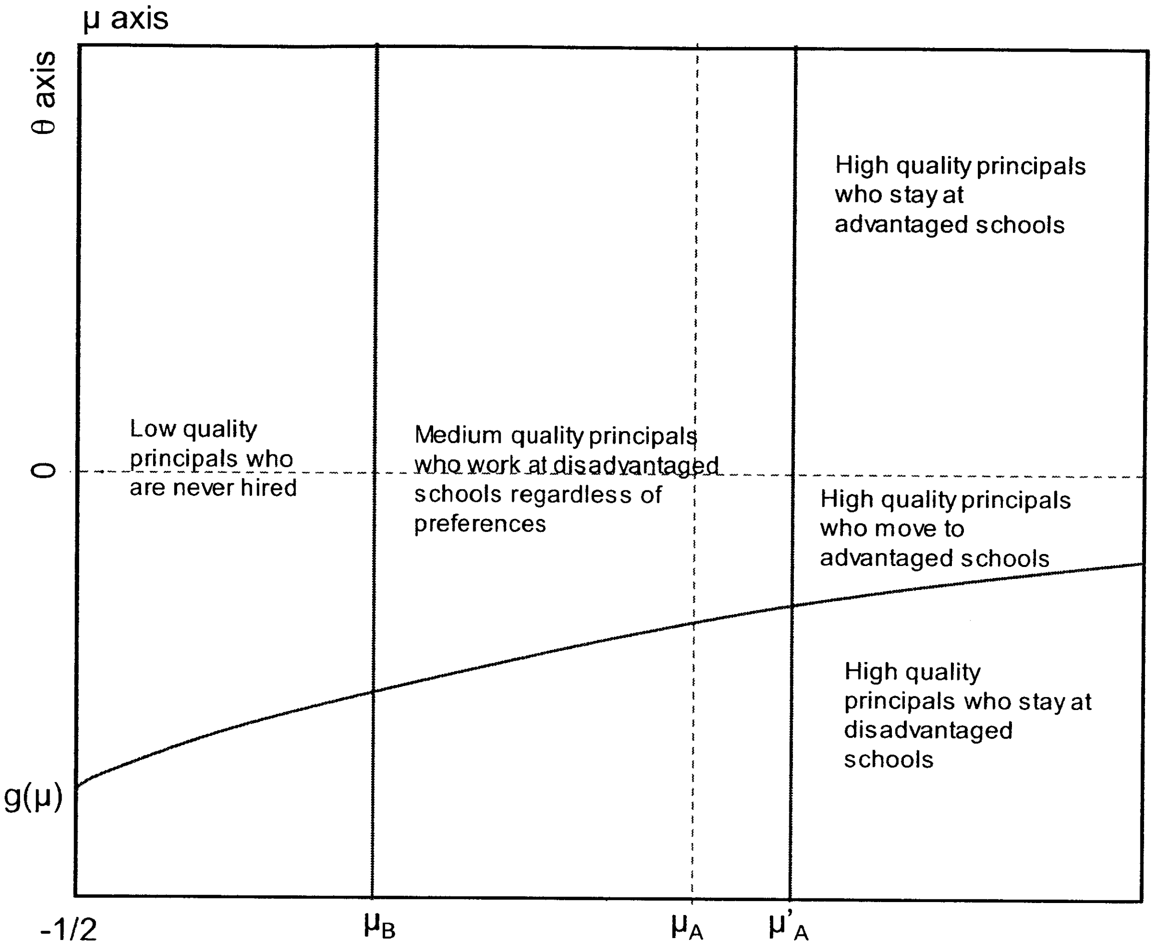
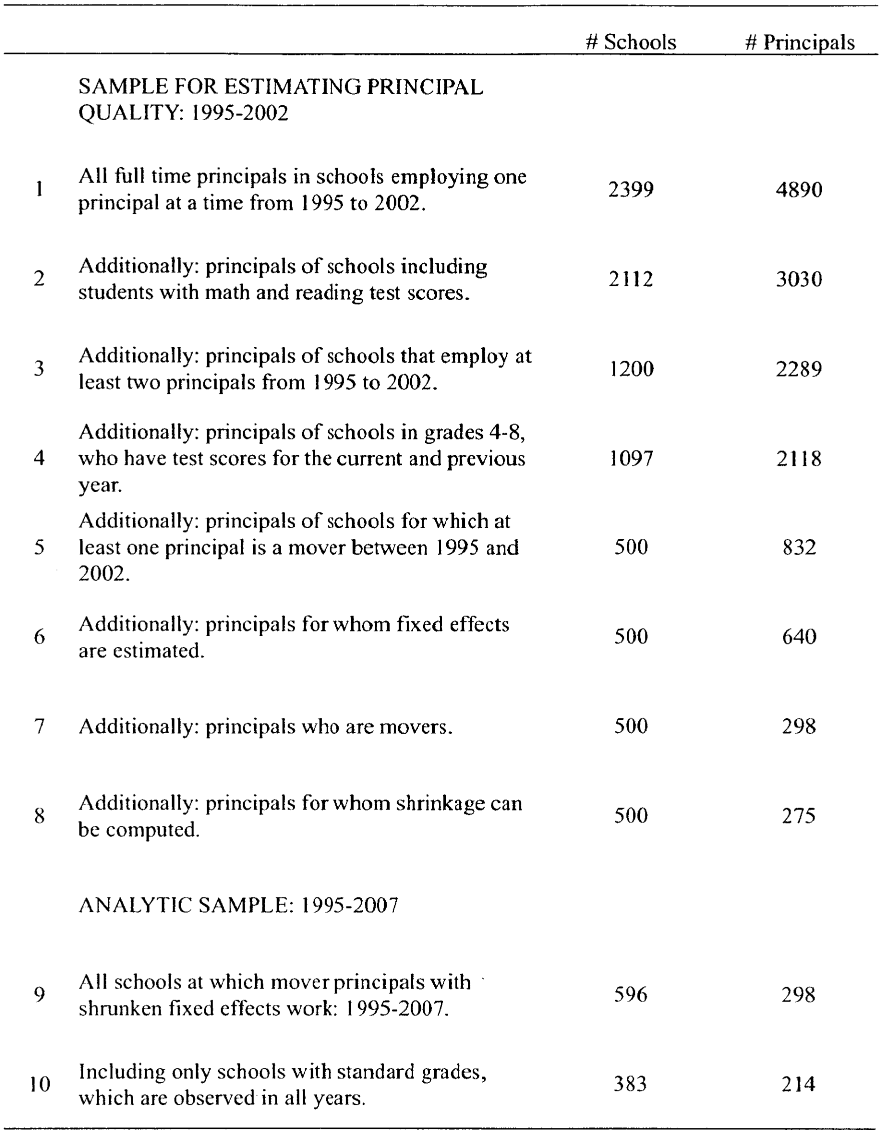

μ axis
θ axis
0
g(μ)
-1/2
μB
μA
μ'A
Low quality principals who are never hired
Medium quality principals who work at disadvantaged schools regardless of preferences
High quality principals who stay at advantaged schools
High quality principals who move to advantaged schools
High quality principals who stay at disadvantaged schools

Figure 2: Allocation of principals after accountability



		# Schools	# Principals
	SAMPLE FOR ESTIMATING PRINCIPAL QUALITY: 1995-2002		
1	All full time principals in schools employing one principal at a time from 1995 to 2002.	2399	4890
2	Additionally: principals of schools including students with math and reading test scores.	2112	3030
3	Additionally: principals of schools that employ at least two principals from 1995 to 2002.	1200	2289
4	Additionally: principals of schools in grades 4-8, who have test scores for the current and previous year.	1097	2118
5	Additionally: principals of schools for which at least one principal is a mover between 1995 and 2002.	500	832
6	Additionally: principals for whom fixed effects are estimated.	500	640
7	Additionally: principals who are movers.	500	298
8	Additionally: principals for whom shrinkage can be computed.	500	275
	ANALYTIC SAMPLE: 1995-2007		
9	All schools at which mover principals with shrunken fixed effects work: 1995-2007.	596	298
10	Including only schools with standard grades, which are observed in all years.	383	214

Figure 3: Sample construction

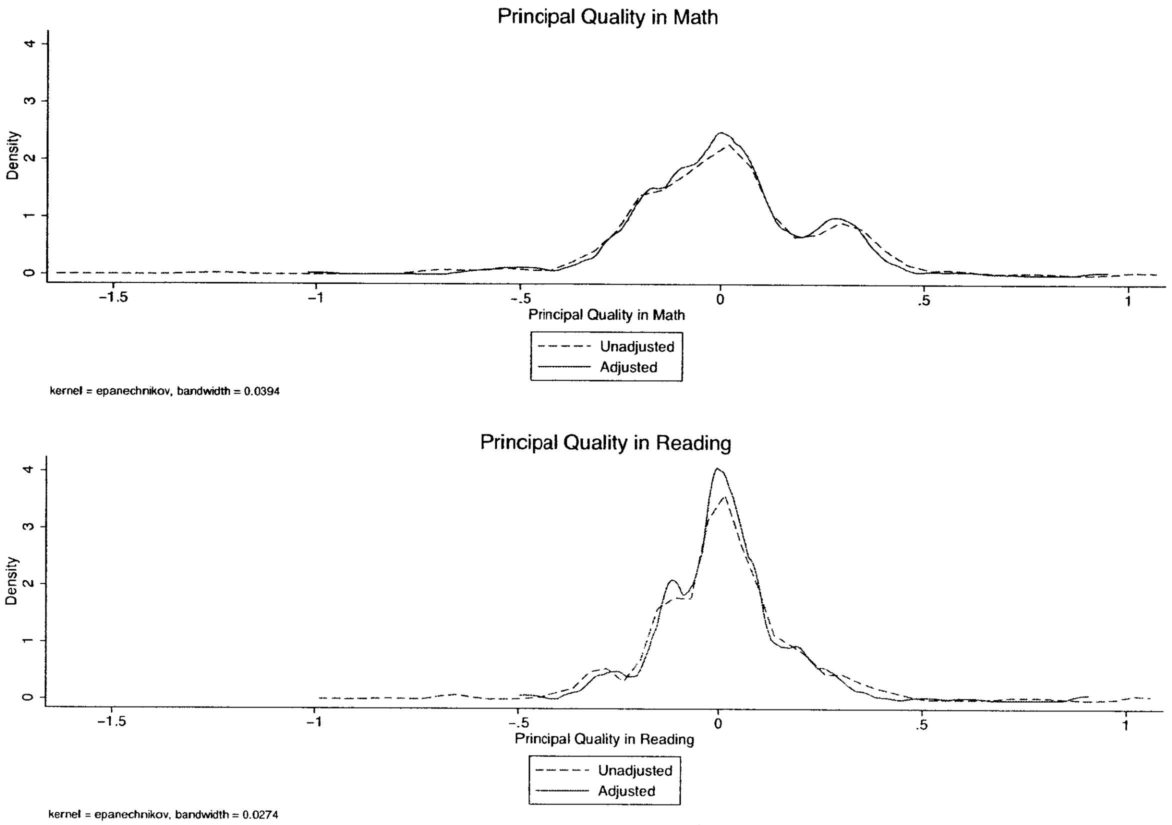
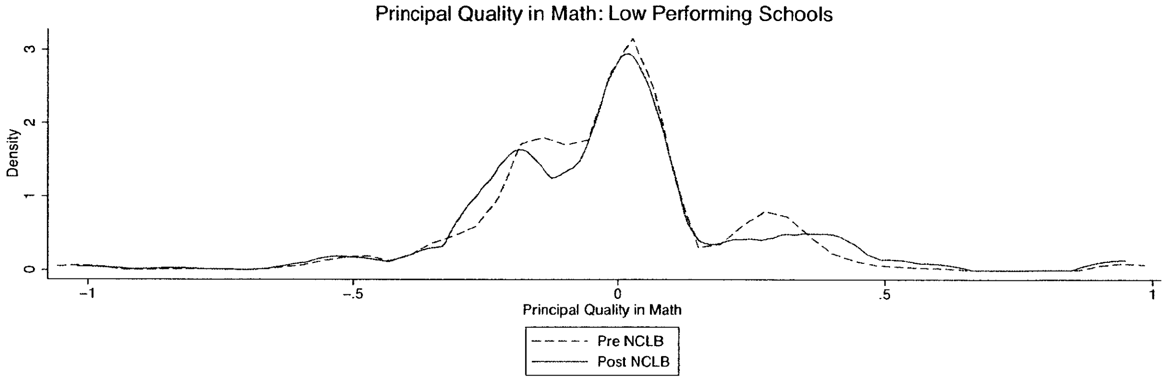
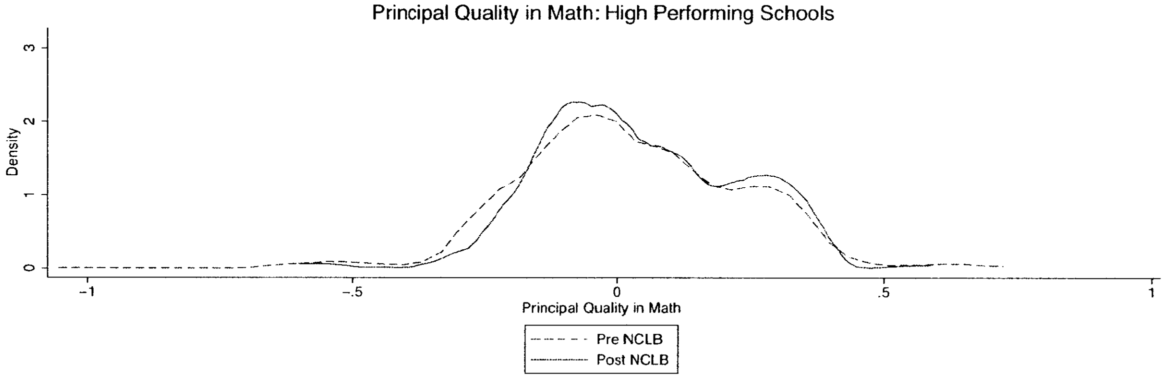


Figure 4: The Distribution of Principal Quality



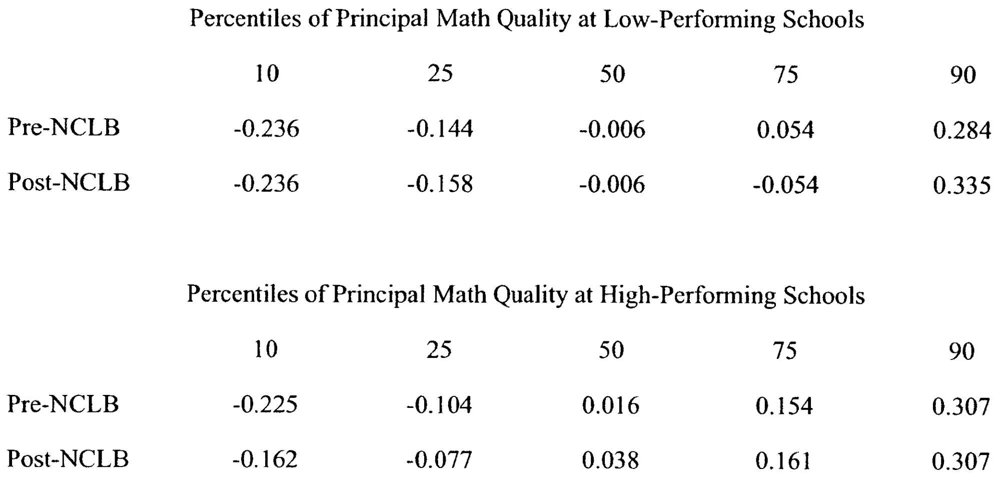
Principal Quality in Math: Low Performing Schools

kernel = epanechnikov, bandwidth = 0.0370



Principal Quality in Math: High Performing Schools

kernel = epanechnikov, bandwidth = 0.0479



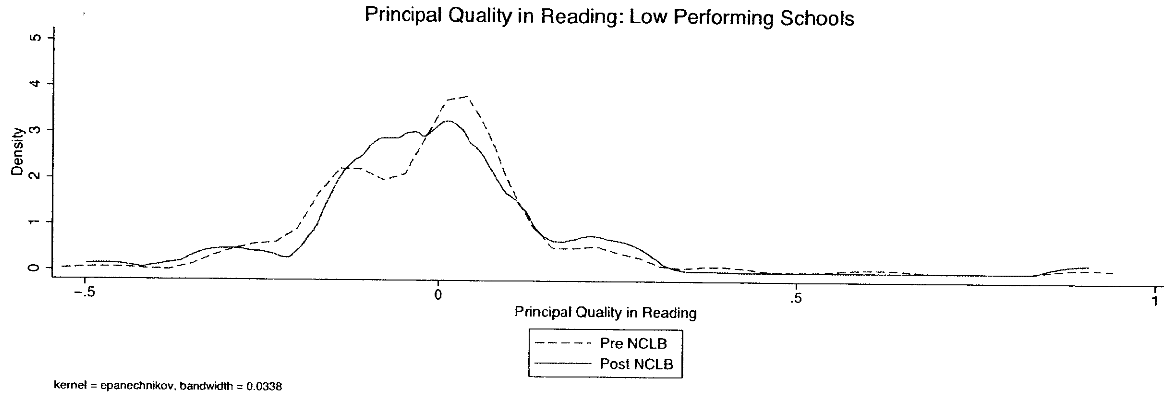
Percentiles of Principal Math Quality at Low-Performing Schools

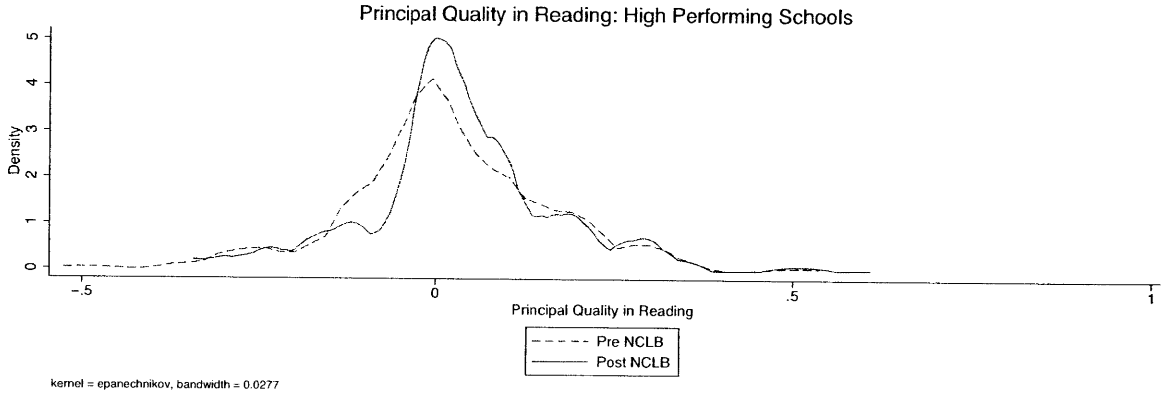
	10	25	50	75	90
Pre-NCLB	-0.236	-0.144	-0.006	0.054	0.284
Post-NCLB	-0.236	-0.158	-0.006	-0.054	0.335

Percentiles of Principal Math Quality at High-Performing Schools

	10	25	50	75	90
Pre-NCLB	-0.225	-0.104	0.016	0.154	0.307
Post-NCLB	-0.162	-0.077	0.038	0.161	0.307

Figure 5: The Distribution of Principal Math Quality Before and After NCLB


Principal Quality in Reading: Low Performing Schools
Density
0
1
2
3
4
5
-.5
0
.5
1
Principal Quality in Reading
Pre NCLB
Post NCLB
kernel = epanechnikov, bandwidth = 0.0338


Principal Quality in Reading: High Performing Schools
Density
0
1
2
3
4
5
-.5
0
.5
1
Principal Quality in Reading
Pre NCLB
Post NCLB
kernel = epanechnikov, bandwidth = 0.0277

Percentiles of Principal Reading Quality at Low-Performing Schools

	10	25	50	75	90
Pre-NCLB	-0.171	-0.118	0.003	0.062	0.171
Post-NCLB	-0.148	-0.098	-0.016	0.064	0.173

Percentiles of Principal Reading Quality at High-Performing Schools

	10	25	50	75	90
Pre-NCLB	-0.143	-0.051	0.007	0.099	0.189
Post-NCLB	-0.143	-0.02	0.018	0.094	0.189

Figure 6: The Distribution of Principal Reading Quality Before and After NCLB

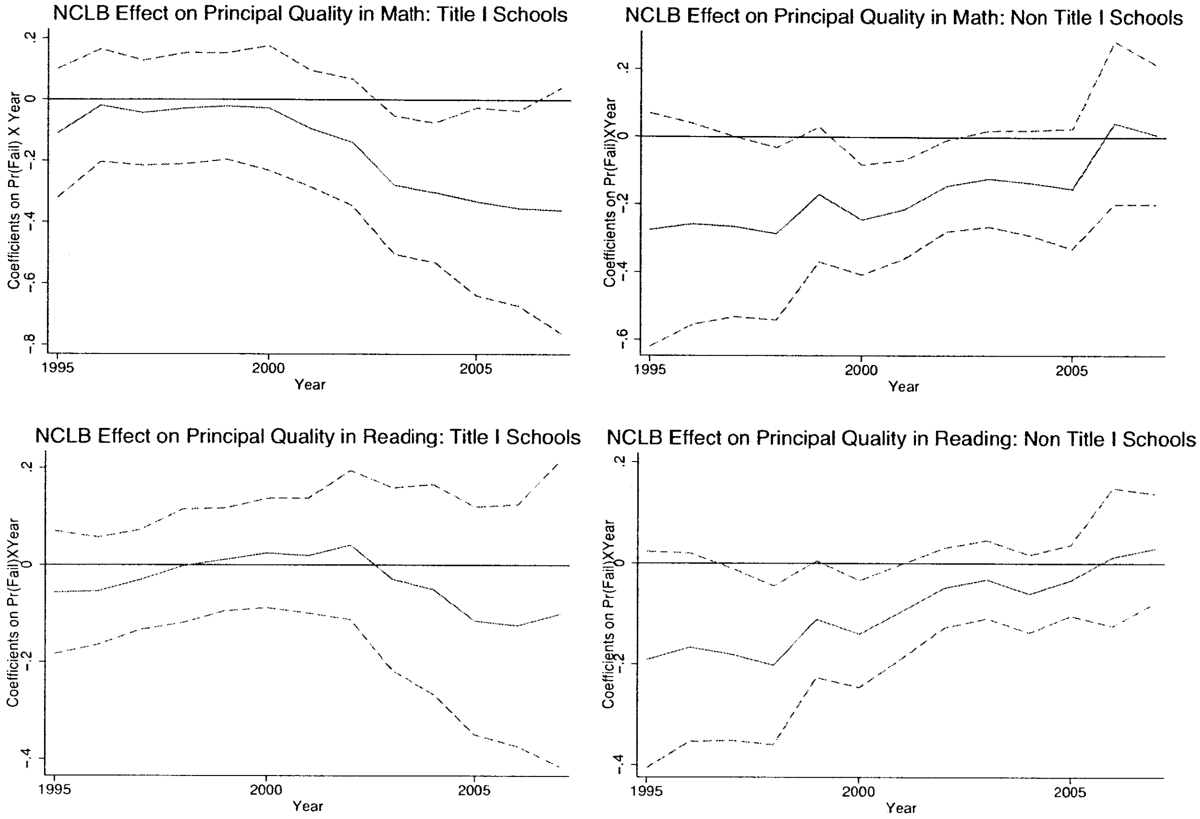


Figure 7: The Timing of NCLB's Effect on Principal Quality

TABLE 1: SUMMARY STATISTICS

	Sample	Universe	P-values
# Schools	383	1605	
# Principals	214	2054	
PRINCIPAL CHARACTERISTICS			
Years in data	10.1	8.4	0.000
	(2.290)	*(2.840)*	
% Ever switch schools	100	65.97	0.000
Years in data at first switch, conditional on switching	3.33	3.39	0.627
	(1.700)	*(2.410)*	
Imputed Age	48.73	48.84	0.268
	(6.365)	*(7.126)*	
Advanced Degree	0.359	0.375	0.437
	(0.480)	*(0.484)*	
State Salary	68,776	69,124	0.000
	(13459)	*(15,706)*	
Principal Tenure (0/1)	0.781	0.814	0.000
	(0.413)	*(0.389)*	
Principal Experience (0/1)	0.958	0.923	0.049
	(0.200)	*(0.267)*	
SCHOOL CHARACTERISTICS			
Pr(Fail) AYP in 2002	0.560	0.466	0.000
	(0.362)	*(0.384)*	
Urban	0.523	0.436	0.022
	(0.500)	*(0.496)*	
Title I	0.620	0.519	0.000
	(0.486)	*(0.500)*	
Proportion Black	0.364	0.318	0.003
	(0.245)	*(0.244)*	
Proportion Hispanic	0.0584	0.0534	0.309
	(0.0662)	*(0.0661)*	
Proportion Asian	0.0212	0.0162	0.006
	(0.0289)	*(0.0249)*	
Proportion White	0.539	0.591	0.000
	(0.261)	*(0.270)*	
Elementary School	0.705	0.533	0.000
	(0.456)	*(0.499)*	
Proportion at Grade Level	0.762	0.765	0.728
	(0.118)	*(0.119)*	
Student-teacher Ratio	14.91	14.94	0.820
	(2.917)	*(2.853)*	
School Size	596.2	658.0	0.000
	(263.5)	*(352.7)*	
Joint			0.000

Notes: Standard devations are in parentheses. Observations are school-year cells. The universe sample include all schools that have 1) one principal at a time; 2) are open in all years between 1995 and 2007; and 3) employ full-time principals who have worked at least one year before 2003. The analytic sample further restricts this sample to schools employing principals who 1) work at least two elementary or middle schools before 2003, and 2) have estimated fixed effects.

TABLE 2: PREDICTORS OF SAMPLE REPRESENTATION AND MOBILITY: ANALYTIC SAMPLE VS. UNIVERSE

	Total years observed	Years observed to date	Switch in next year	Retire in next year
	(1)	(2)	(3)	(4)
PANEL 1: PREDICTED PROBABILITY OF FAILING AYP				
Pr(Fail) AYP in 2002 X 1(Sample Principal)	-0.111	0.218	-0.032	-0.008
	(0.342)	*(0.340)*	*(0.022)*	*(0.014)*
Pr(Fail) AYP in 2002	-0.235	-0.335***	0.028***	-0.006
	(0.160)	*(0.107)*	*(0.008)*	*(0.005)*
1(Sample Principal)	1.456***	0.640***	0.077***	-0.019**
	(0.231)	*(0.210)*	*(0.015)*	*(0.008)*
PANEL 2: PROPORTION OF STUDENTS WHO ARE WHITE				
Proportion White X 1(Sample Principal)	-0.361	-0.553	0.018	0.010
	(0.471)	*(0.427)*	*(0.029)*	*(0.016)*
Proportion White	0.678***	-0.423***	-0.048***	-0.068***
	(0.211)	*(0.127)*	*(0.010)*	*(0.006)*
1(Sample Principal)	1.668***	0.990***	0.048***	-0.032***
	(0.309)	*(0.247)*	*(0.018)*	*(0.009)*
PANEL 3: PERCENT OF STUDENTS WHO ARE ON FREE OR REDUCED LUNCH				
Free/Reduced Lunch X 1(Sample Principal)	1.065**	0.159	0.045	-0.054**
	(0.489)	*(0.550)*	*(0.037)*	*(0.025)*
Proportion Free/Reduced Lunch	-1.287***	0.086	0.012	0.091***
	(0.245)	*(0.165)*	*(0.014)*	*(0.010)*
1(Sample Principal)	1.059***	0.641***	0.042***	-0.006
	(0.238)	*(0.231)*	*(0.016)*	*(0.010)*

Notes: Standard errors are in parenthesis. Each column in each panel is its own separate regression. Observations are school-year cells. The universe sample include all schools that have 1) one principal at a time; 2) are open in all years between 1995 and 2007; and 3) employ full-time principals who have worked at least one year before 2003. The analytic sample further restricts this sample to schools employing principals who 1) work at least two elementary or middle schools before 2003, and 2) have estimated fixed effects. Total years refers to the total number of years a principal is observed in the NC state data, including years where she is employed by a school that is not always open or employs more than one principal. Years in observed to date, switching, and retirement are all defined on this extended sample.

TABLE 3: HOW IS PR(FAIL) CORRELATED WITH SCHOOL CHARACTERISTICS?

	Sample	Universe
Mean of Pr(Fail):	0.549	0.461
% Black	0.839*** *(0.0874)*	0.647*** *(0.0555)*
% Hispanic	1.137*** *(0.190)*	1.222*** *(0.116)*
% Asian	1.041** *(0.451)*	0.375 *(0.288)*
% Other	0.692*** *(0.155)*	0.463*** *(0.110)*
% Free Lunch	0.322*** *(0.120)*	0.595*** *(0.0769)*
Students (1000s)	0.150** *(0.0749)*	0.0363 *(0.0335)*
Elementary School	-0.173*** *(0.0381)*	-0.129*** *(0.0159)*
N	1791	19221
R2	0.601	0.518

Notes: Pr(Fail) is the probability that a school fails AYP based on demographics, urbanicity, and school level from 1995 to 2002. This table does NOT report results from the actual estimation of Pr(Fail), which involves linear demographics for each year interacted with an elemenatary school dummy as well as yearly quadratics and cubics in student demographics. See text for details.

TABLE 4: CORRELATES OF PRINCIPAL QUALITY BEFORE AND AFTER NCLB

Dep. Var.	Math FE		Reading FE	
	Pre	Post	Pre	Post
	(1)	(2)	(1)	(2)
Pr(Fail)	-0.118** *(0.048)*	-0.171*** *(0.064)*	-0.076** *(0.032)*	-0.156** *(0.067)*
% At grade level	0.456*** *(0.156)*	0.577*** *(0.200)*	0.105 *(0.107)*	0.619*** *(0.204)*
% White	0.077 *(0.080)*	0.264** *(0.126)*	-0.030 *(0.064)*	0.289** *(0.129)*
% Black	-0.094 *(0.083)*	-0.306** *(0.140)*	0.051 *(0.069)*	-0.338** *(0.143)*
% Hispanic	0.204 *(0.317)*	-0.117 *(0.421)*	-0.120 *(0.196)*	-0.122 *(0.430)*
% Free Lunch	-0.059 *(0.090)*	-0.248** *(0.110)*	0.004 *(0.062)*	-0.254** *(0.115)*

Notes: Each cell is a separate regression of the indicated variable on estimates of principal value-added, controlling for district fixed effects only, weighted by the inverse variance of the principal quality measure. Sample is the analytic sample of principals with estimated fixed effects. Standard errors are clustered at the school level.

TABLE 5: EFFECTS OF NCLB ON THE ALLOCATION OF PRINCIPAL QUALITY

	Principal Math Quality		Principal Reading Quality	
	Title I	Non Title I	Title I	Non Title I
	(1)	(2)	(3)	(4)
Pr(Fail) X 1(Year>2002)	-0.206***	0.087	-0.083	0.069*
	(0.075)	*(0.067)*	*(0.069)*	*(0.037)*
Pr(Fail)	-0.070	-0.206***	-0.000	-0.113**
	(0.078)	*(0.074)*	*(0.046)*	*(0.045)*
Observations	1120	671	1117	665
R-squared	0.442	0.377	0.503	0.485

Notes: Sample is the set of schools that 1) employ one principal at a time; 2) are open in all years between 1995 and 2007; and 3) employ full-time principals. Principals included must 1) work at least two elementary or middle schools before 2003, and 2) have estimated fixed effects. Pr(Fail) is the probability that a school will fail AYP based on prior demographics, urbanicity, and school level. 1(Year>2002) is a dummy for post 2002. Regressions control for district and year fixed effects, linear time trends for each district, and principal age and age squared. Standard errors are clustered at the school level. Regressions are weighted by the inverse variance of the relevant principal fixed effect.

TABLE 6: EFFECT OF NCLB ON THE ALLOCATION OF PRINCIPAL MATH QUALITY AT TITLE I SCHOOLS , BY DISTRICT CHARACTERISTICS

	Principal Math Quality			
	Large Districts	Small Districts	Urban	Non-Urban
	(1)	(2)	(3)	(4)
Pr(Fail) X 1(Year>2002)	-0.205*	-0.116	-0.235	-0.053
	(0.104)	*(0.109)*	*(0.000)*	*(0.093)*
Pr(Fail)	-0.120	-0.016	-0.064	-0.337**
	(0.111)	*(0.089)*	*(0.000)*	*(0.164)*
Observations	1791	1120	671	1782
R-squared	0.317	0.442	0.377	0.376

Notes: Sample is the set of schools that 1) employ one principal at a time; 2) are open in all years between 1995 and 2007; and 3) employ full-time principals. Principals included must 1) work at least two elementary or middle schools before 2003, and 2) have estimated fixed effects. Pr(Fail) is the probability that a school will fail AYP based on prior demographics, urbanicity, and school level. 1(Year>2002) is a dummy for post 2002. Regressions control for district and year fixed effects, linear time trends for each district, and principal age and age squared. Standard errors are bootstrapped and clustered at the school level. Regressions are weighted by the inverse variance of the relevant principal fixed effect.

TABLE 7: EFFECT OF NCLB ON AGGREGATE PRINCIPAL MOBILITY

	Sample - Switch	Universe - Switch	Universe - Retire
	(1)	(2)	(3)
Mean of dep. var.	0.153	0.096	0.089
Pr(Fail) X 1(Year>2002)	0.080	0.017	-0.003
	(0.058)	*(0.015)*	*(0.016)*
Pr(Fail)	-0.043	0.014	0.009
	(0.035)	*(0.010)*	*(0.007)*
Observations	1714	13447	13447
R-squared	0.119	0.039	0.106

Notes: The universe sample include all schools that have 1) one principal at a time; 2) are open in all years between 1995 and 2007; and 3) employ full-time principals who work at least one year prior to 2003. The analytic sample further restricts this sample to schools employing principals who 1) work at least two elementary or middle schools before 2003, and 2) have estimated fixed effects. Switch is an indicator for whether a principal becomes a principal at a different school in the following year. Retire is a dummy equal to one if the principal is no longer working as a principal in the following year. Pr(Fail) is the probability that a school will fail AYP based on prior demographics, urbanicity, and school level. 1(Year>2002) is a dummy for post 2002. Regressions control for district and year fixed effects, linear time trends for each district, and principal age and age squared. Standard errors are clustered at the school level.

TABLE 8: EFFECT OF NCLB ON PRINCIPAL-SCHOOL MATCHING, BY PRINCIPAL MATH QUALITY

		Characteristics of the school to which a principal moves, conditional on moving			
	Switch	Pr(Fail)	% At Grade Level	% White	Title I
	(1)	(2)	(3)	(4)	(5)
Mean of Dep. Var.	0.153	0.555	0.755	0.536	0.537
Pr(Fail) X 1(Year>2002) for Math Quality < Median	0.106	-0.190	-0.112	-0.243	0.285
	(0.082)	*(0.326)*	*(0.099)*	*(0.171)*	*(0.361)*
Pr(Fail) X 1(Year>2002) for Math Quality > Median	0.056	-0.986**	0.058	0.537***	-0.681
	(0.073)	*(0.478)*	*(0.087)*	*(0.197)*	*(0.457)*
Observations	1714	220	241	263	261
R-squared	0.119	0.690	0.661	0.666	0.587

Notes: Reported are coefficients on Pr(fail)X1(Year>2002) for each ability group. Sample is the set of schools that 1) employ one principal at a time; 2) are open in all years between 1995 and 2007; and 3) employ full-time principals. Principals included must 1) work at least two elementary or middle schools before 2003, and 2) have estimated fixed effects. Pr(Fail) probability of failing AYP based on prior demographics, urbanicity, and school level. 1(Year>2002) is a dummy for post 2002. 1(Year>2002) is a dummy for post 2002. Regressions control for district and year fixed effects, linear time trends for each district, and principal age and age squared. All regressions include year and district fixed effects and district by year linear trends. Standard errors are clustered at the school level.

TABLE 9: EFFECT OF NCLB ON PRINCIPAL-SCHOOL ASSIGNMENT BASED ON NAIVE ABILITY

PANEL 1: DISTRIBUTION OF QUALITY

	Principal Math Quality	
	Title I	Non Title I
Dep. Var.	(1)	(2)
Pr(Fail) X 1(Year>2002)	-0.017	-0.017
	(0.045)	*(0.048)*
Pr(Fail)	-0.046	-0.187***
	-0.037	*(0.053)*
Observations	1240	766
R-squared	0.463	0.489

PANEL 2: MOBILITY

		Characteristics of the school a principal moves to, conditional on moving		
	Switch (0/1)	Pr(Fail)	% At Grade Level	% White
	(4)	(5)	(6)	(7)
Pr(Fail) X 1(Year>2002) for Math Quality < Median	0.176**	-0.461*	0.074	-0.064
	(0.078)	*(0.256)*	*(0.102)*	*(0.212)*
Pr(Fail) X 1(Year>2002) for Math Quality > Median	0.004	-0.424	0.061	0;162
	(0.070)	*(0.389)*	*(0.066)*	*(0.215)*
Observations	1920	250	272	298
R-squared	0.116	0.619	0.610	0.647

APPENDIX TABLE A: EFFECT OF NCLB ON DISTRICT SALARY SUPPLEMENTS

	Sample		Universe	
	% Receiving	Avg. supplement (1000s)	% Receiving	Avg. supplement (1000s)
	(1)	(2)	(3)	(4)
Mean of dep. var.	0.968	7.031	0.929	5.841
Pr(Fail) X 1(Year>2002)	-0.043	-3.039	-0.000	-2.672
	(0.030)	*(2.252)*	*(0.044)*	*(1.764)*
Pr(Fail)	0.010	9.875*	0.072	5.634
	(0.037)	*(5.216)*	*(0.069)*	*(4.344)*
Observations	323	323	672	672
R-squared	0.013	0.069	0.014	0.015

Notes: Regression is at the district-year level for the years 2002-2007. % Receiving indicates the percentage of principals receiving a district supplement, average supplement includes zeros, in 2007 dollars. Year fixed effects are included. Standard errors are clustered at the district level and the regression is weighted by district size.

APPENDIX TABLE B: DIFFERENTIAL CHANGES IN SAMPLE EXIT BY SCHOOL CHARACTERISTICS

	RHS School Characteristics			
Dep. Var.: 1(School leaves sample in the next year)	Pr(Fail)	% Grade Level	% White	% Free Lunch
	(1)	(2)	(3)	(4)
FULL SAMPLE				
School Characteristic X 1(Year>2002)	0.050	-0.219	0.112	0.047
	(0.075)	*(0.306)*	*(0.106)*	*(0.107)*
School Characteristic	-0.042	-0.121	-0.093	0.008
	(0.037)	*(0.135)*	*(0.065)*	*(0.062)*
Observations	1714	1502	1714	1707
R-squared	0.125	0.127	0.125	0.123
TITLE I ONLY				
School Characteristic X 1(Year>2002)	-0.012	-0.068	0.164	-0.002
	(0.121)	*(0.364)*	*(0.160)*	*(0.151)*
School Characteristic	0.098	-0.231	-0.250**	0.110
	(0.061)	*(0.201)*	*(0.102)*	*(0.094)*
Observations	1075	939	1075	1068
R-squared	0.168	0.168	0.171	0.167

Notes: Sample is the set of schools that 1) employ one principal at a time; 2) are open in all years between 1995 and 2007; and 3) employ full-time principals. Principals included must 1) work at least two elementary or middle schools before 2003, and 2) have estimated fixed effects. Pr(Fail) is the probability that a school will fail AYP based on prior demographics, urbanicity, and school level. 1(Year>2002) is a dummy for post 2002. Regressions include year fixed and district fixed effects, principal age and age squared, and district by year linear trends.

APPENDIX TABLE C: EFFECT OF NCLB ON THE QUALITY OF THE NEXT PRINCIPAL WORKING AT A SCHOOL, RESTRICTED SAMPLE

	Math Quality		Reading Quality	
	Title I	Non Title I	Title I	Non Title I
	(1)	(2)	(3)	(4)
Pr(Fail) X 1(Year>2002)	-0.864	-0.426	−1.380***	-0.321
	(0.968)	*(2.598)*	*(0.489)*	*(1.490)*
Pr(Fail)	0.398	1.689	0.409	0.467
	(0.715)	*(1.091)*	*(0.364)*	*(0.739)*
Observations	56	34	56	34
R-squared	0.789	0.949	0.704	0.826

Notes: Sample includes only school-year observations for which a school changes principals in the next year in which it is observed, and for which I observe the new principal's estimated fixed effect. The dependant variable is the quality of the next principal.

APPENDIX TABLE D: EFFECTS OF NCLB ON THE ALLOCATION OF PRINCIPAL QUALITY WITH BOOTSTRAPPED STANDARD ERRORS

	Principal Math Quality		Principal Reading Quality	
	Title I	Non Title I	Title I	Non Title I
	(1)	(2)	(3)	(4)
Pr(Fail) X 1(Year>2002)	-0.206**	0.087	-0.083	0.069**
	(0.092)	*(0.061)*	*(0.068)*	*(0.034)*
Pr(Fail)	-0.070	-0.206***	-0.000	-0.113**
	(0.087)	*(0.069)*	*(0.050)*	*(0.045)*
Observations	1120	671	1117	665
R-squared	0.442	0.377	0.503	0.485

Notes: Sample is the set of schools that 1) employ one principal at a time; 2) are open in all years between 1995 and 2007; and 3) employ full-time principals. Principals included must 1) work at least two elementary or middle schools before 2003, and 2) have estimated fixed effects. Pr(Fail) is the probability that a school will fail AYP based on prior demographics, urbanicity, and school level. 1(Year>2002) is a dummy for post 2002. Regressions control for district and year fixed effects, linear time trends for each district, and principal age and age squared. Standard errors are clustered at the school level. Regressions are weighted by the inverse variance of the relevant principal fixed effect.

APPENDIX TABLE E: EFFECT OF NCLB ON THE ALLOCATION OF PRINCIPAL MATH QUALITY AT TITLE I SCHOOLS WITH ALTERNATIVE SPECIFICATIONS

	Alternative Principal Quality Measure: Using lagged test scores to measure FE		Alternative Failure Probability Measure: using predicted % of targets failed		Alternative Principal Quality Measure: using unshrunken fixed effects	
	Title I	Non Title I	Title I	Non Title I	Title I	Non Title I
Math FE	(1)	(2)	(3)	(4)	(5)	(6)
Pr(Fail) X 1(Year>2002)	-0.073	0.011	-0.276**	-0.049	-0.213**	0.096
	(0.057)	*(0.046)*	*(0.138)*	*(0.080)*	*(0.083)*	*(0.075)*
Pr(Fail)	-0.095	-0.090	0.041	-0.109	-0.070	-0.223***
	(0.059)	*(0.057)*	*(0.091)*	*(0.095)*	*(0.086)*	*(0.084)*
Observations	1109	676	1120	671	1120	671
R-squared	0.508	0.311	0.442	0.348	0.445	0.338

Notes: Sample includes observations for all schools employing principals who were movers prior to 2003, and for whom I have estimated fixed effects. Sample splits are based on Title I status in 2002-03. Regressions in the first panel use math fixed effects computed with lagged student test scores. School performance is defined as the probability that a school will fail AYP based on pre-period demographics. Regressions in the second panel are based on school performance measured as percent of AYP targets a school is predicted to fail based on pre-period characteristics. Regressions in the final panel use unadjusted fixed effects. All regressions control for district and year fixed effects, linear time trends for each district, and principal age and age squared. Standard errors are clustered at the school-year level. Regressions are weighted by the inverse variance of the relevant principal fixed effect. Standard errors are clustered at the school level.

Chapter 4

Cheaper by the Dozen: Using Sibling Discounts at Catholic Schools to Estimate the Price Elasticity of Private School Attendance[1]

4.1 Introduction

One of the most important public policy debates over the past decade has been the appropriate role of school choice in U.S. education policy. Starting with Milton Friedman (1962), proponents of school choice have advocated funding schools through a system of portable vouchers that would allow families to purchase educa-

[1] This chapter is coauthored with Susan Dynarski and Jonathan Gruber

tion at the school of their choice and, thereby, create competitive pressures on public schools. The generation of such competitive pressures depends upon the willingness and ability of parents to move their children between schools, particularly from public schools to private schools. In particular, the response to a voucher program depends critically on the price elasticity of demand for private schooling.

Evidence of the responsiveness of families to private schooling prices is remarkably thin. Derek Neal (2002) notes that there is extensive research on the effect of private schools on student outcomes, but comparatively little empirical evidence of how tuition prices affect the decision to attend a private school.[2] This is a challenging parameter to estimate. In fact, a regression of quantity on price is frequently used in econometrics textbooks to illustrate the challenges of estimating causal parameters with observational data. Without (quasi-) random variation in the tuition prices set by schools, a regression of quantity on price captures movement along both the supply curve and the demand curve.

We exploit a unique source of variation in tuition prices to estimate the price elasticity of demand for private schooling. The majority of Catholic elementary schools offer sibling discounts. These discounts reduce schooling costs for families that, in a given year, enroll more than one child in a single Catholic school. We have collected data on these discounts from schools representing over half of Catholic school enrollment in the US. On average, the tuition charged for the second sibling enrolled in a Catholic elementary school is 25 percent lower than tuition for the first sibling, and the tuition is 36 percent lower for the third sibling than for the first sibling. Each school establishes its own pricing schedule; about half of schools offer

[2] In order to predict how a voucher program would affect the distribution of students between public and private schools, Neal relies on theoretical predictions of Epple and Romano (1998, 2002). The present paper aims to provide an empirical foundation for this critical prediction.

the discounts. Discounts vary dramatically, even within a metropolitan area. As a result of these pricing schedules, the tuition prices faced by a family are a function of the interaction of the number and spacing of their children with the pricing policies of the local Catholic school.

To execute this strategy, we have collected a new data set of the tuition schedules offered at Catholic schools. We have collected, from 60 Catholic dioceses, information on the tuition schedules of 1760 schools, representing over one-third of all Catholic school enrollment in the U.S. We match this newly collected tuition data to restricted-use Census data that identify the block in which a household is located. This variation in tuition prices across families within a given neighborhood allows us to include in our demand equation a detailed set of block-group fixed effects to control for any unobserved determinants of demand that vary across space. Since the discounts vary considerably across schools, we are also able to control flexibly for the number, spacing and ages of children in each family, thereby absorbing any nationwide, underlying relationship between family composition and private school attendance.

We find that a standard deviation decrease in tuition prices increases the probability that a family will send its children to private school by over two thirds of a percentage point. This translates into an elasticity of Catholic school attendance with respect to tuition costs of -0.19. This average effect masks substantial heterogeneity in the response to price, with lower income families and those with less educated parents being more price sensitive. These results strongly reject the assumption made in previous studies (e.g., Figlio and Stone, 2001; Lankford and Wyckoff, 2001) that the students that vouchers would induce into private school would look demographically similar to current private school students.

Our paper proceeds as follows. In Section 2, we present a simple theoretical motivation for our topic, drawing on the seminal model of Peltzman (1973) to highlight the centrality of the price elasticity of demand in the evaluation of voucher programs. In Section 3, we discuss empirical challenges in the estimation of the price elasticity of demand for private schooling and critically review the existing literature on this topic. Section 4 lays out our identification strategy and discusses the characteristics of the tuition and census data. Our basic results and sensitivity tests are presented in Section 5, while Section 6 presents evidence on the heterogeneity of price elasticity and Section 7 concludes.

4.2 Theoretical Motivation

The starting point for our analysis is the seminal school choice model of Peltzman (1973). Figure 1A illustrates a familys choice between education and all other goods in the absence of the public provision of education. The family has a total budget of Q_0; the price of private education is p (the slope of the budget line), and the price of other goods is normalized to 1. There is a smooth tradeoff between the consumption of education and of other goods. The optimal choice is E_1, the point at which (constrained by budget Q_0) a consumers marginal rate of substitution between other goods and education is p. Now introduce the public provision of education. The public sector provides education of amount E_F. Consumers can spend Q_0 on other goods and still consume E_F. Parents who wish to purchase a higher quantity must send their children to private school, thereby forgoing their entitlement to free public education (there is no "topping off" allowed). Consider the schooling choice of a family with an indifference curve tangent to the budget constraint at X_1. This family could choose private schooling, and obtain more education at $E_1 > E_F$ but consumption would fall. Given this familys marginal rate of substitution between other goods and education, it would prefer the free, public education.

Next we add vouchers to the model. The private enrollment response to a voucher depends on the size of the voucher and the preferences of consumers. A larger voucher (Figure 1C) will move more families into private school than a smaller voucher (Figure 1D). Were all families identical, all families would choose the same schooling option (e.g. private school in Figure 1C, public in Figure 1D). But if preferences are heterogeneous, a large voucher (Figure 1E) will move into private school some families unmoved by a small voucher (Figure 1F). Heterogeneity in underlying tastes can lead to a smooth aggregate relationship between voucher levels and choice of private schooling.

Thus, we can derive a demand curve for the relationship between the cost of going to private school rather than public school and the share of families choosing private rather than public school. This demand curve will depend on the relative densities of different types in the population. In principle, the demand elasticity could be very non-linear, because different price changes could hit individuals with very different marginal rates of substitution across public and private schooling.

4.3 Empirical Issues in Estimating the Price Elasticity of Private Schooling

In principle, estimating the price elasticity of demand for private schooling is straightforward: individual school enrollment (y) is regressed on the price of nearby private schools. In Equation (4.1), a binary measure of the private school attendance of child i living in family j living in area b is regressed on the price of private schools in geographic area b:

$$y_{ijb} = \beta_0 + \beta_1 \text{price}_b + \varepsilon_{ijb} \tag{4.1}$$

Equation (4.1) estimates the association between a dollar increase in private school tuition and the probability that individual i attends private school. We would like to interpret $\hat{\beta}_1$ as the causal impact of price on private school attendance. However, the price faced by individual i is likely a function of omitted variables correlated with the demand for private schooling. For example, in high-income areas both price and the attendance rate will be above average if private school is a normal good. Controlling for income, along with other covariates, is one way to deal with this problem:

$$y_{ijb} = \beta_0 + \beta_1 \text{price}_b + \beta_2 X_j + \beta_3 X_b + \varepsilon_{ijb} \tag{4.2}$$

Control variables typically included in this type of regression include characteristics of the parents in family j (e.g., marital status, race, education, age) and characteristics of the geographic area b (e.g., poverty rate, population density, local public school characteristics). The central weakness in this approach is that cross-sectional variation in equilibrium tuition prices reflects not only variation in the supply of schools (useful for the purposes of identifying a demand elasticity) but also in the demand for schools (which will bias estimates of demand elasticities).

A number of studies have taken the empirical approach of Equation (4.2). Keeler and Kriesel (1994) estimate the relationship between tuition prices and the share of children in the district attending private schools in 105 school districts in Georgia; their cross-sectional estimates suggest an elasticity of -1.04. Buddin, Cordes and Kirby (1998) estimate the relationship between tuition prices and private school attendance in California and conclude that "[T]he propensity of families to choose private schools is insensitive to out-of-pocket tuition costs, which implies that providing school vouchers would encourage few families to shift from public to private schools." Erekson (1982) examines the relationship between private school attendance and Catholic school prices in New York State and finds that Catholic school attendance increases with tuition prices. Lankford and Wyckoff (1992), using similar data, find a negative relationship between price and attendance. Chiswick and Koutromanes (1996) correlate private school attendance with variation in private school tuition at the state level. They estimate that an increase in tuition prices from $1,000 to $4,000 decreases the probability of choosing a private school from 23 percent to 17 percent, calculating an overall price elasticity of -0.48.

Long and Toma (1988) model the determinants of private school attendance using 1970 and 1980 Census data. They are primarily interested in the relationship between race, income and private school attendance, but also include a "tuition" variable in their models. Since they do not have direct measures of private school tuition they proxy for tuition costs with the state-level average of private school salary costs per employee. They estimate the relationship between this tuition proxy and private school attendance for several different years and for different levels of schooling, obtaining estimates that range from statistically insignificant and negative to statistically significant and positive.

The mixed and sometimes perverse results in this literature may reflect a common flaw: a lack of exogenous variation in private school prices. The prices of private school are equilibrium outcomes, determined by both the supply of and demand for private schooling. The price coefficient in Equation (4.2) is identified only under the very strong assumption that observable characteristics fully capture variation in the demand for private schooling. But some determinants of demand are unobservable in typical datasets and so cannot be included in this vector of controls. In areas with a high taste for private school, we will observe both high enrollment rates and high tuitions, which both reflect the higher demand for private school. This will positively bias our price coefficient. Alternatively, in the case of Catholic schools, the bias could be negative: the degree of commitment to the Catholic faith is likely positively correlated with demand for Catholic schooling and negatively correlated with price, since committed parishioners will subsidize the local school's tuition costs with their donations. This would cause us to overestimate the price elasticity of demand for private school attendance.

Ideally, we would randomly assign private school vouchers of varying values, observe responses, and thereby estimate the price elasticity of demand for private schooling. In fact, the randomized assignment of vouchers has occurred in Milwaukee, New York, Dayton, and Washington, DC. Analysts of these experiments are primarily concerned with estimating the impact of private school attendance on student performance (Rouse, 1998; Witte and Thorn, 1996; Mayer et al., 2002). They typically use instrumental variables regression in their empirical analysis, with the reaction of families schooling choices to the offer of a voucher forming the first-stage and the effect of private school attendance on educational outcomes forming the second stage.

These studies do not calculate the demand elasticities implied by the first stage. This is understandable, since their central identification concern is the exogeneity of the relationship between the voucher offer and schooling choices, not the size of the relationship. The price changes are large and discrete, and so are the changes in attendance rates, so the calculated elasticities are very sensitive to whether the baseline attendance rate used in the calculation is that of the treatment or the control group. If we choose the control group as the "base," the elasticities are -6 in Dayton, -9 in Washington and -23 in New York City, while if the treated group is the base case then the elasticities are -1.4, -1.3 and -1.4, respectively. Either approach suggests that families are highly elastic in their response to tuition price. However, a plausible explanation for these magnitudes is that the studied population is non-representative: only those who most desire to attend private school, and whose enrollment hinges on the provision of a subsidy, may be willing to submit to the time demands of a randomized trial over several years (e.g., meetings at nights and on weekend, standardized testing, lengthy surveys that ask for personal information). The bottom line is that the voucher trials produce internally valid estimates of the price elasticity of private school attendance but they are of limited external validity.

A related body of research seeks to predict who will be shifted into private schools by a voucher by describing the population of students who currently attend private schools. These studies assume that the type of student that currently attends private school is the type of student that will be induced into private school by a voucher. This is a strong assumption, one that our empirical research can test directly. Figlio and Stone (2001) use NELS data to show that private schools are disproportionately attended by white students whose parents are of high socioeconomic status. Lankford and Wyckoff (2001) examine the relationship between students and family characteristics and school choice, again using NELS. They find that higher income families are more likely to send their children to private schools.

4.4 Identification Strategy

The key threat to the internal validity of the observational studies we have discussed is that tuition prices charged by private schools are plausibly driven by local demand for private schooling. We address this threat by controlling for an extremely fine set of neighborhood fixed effects:

$$y_{jb} = \beta_0 + \beta_1 \text{price}_{jb} + \delta_b + \varepsilon_{jb} \tag{4.3}$$

In this equation, y_{jb} indicates the private school choice of family j who lives in neighborhood b. Our key explanatory variable is price$_{jb}$, the tuition charged by the private school located nearest the neighborhood. δ_b denotes a set of neighborhood fixed effects. The neighborhood fixed effects absorb any variation between neighborhoods in the unobserved and unobserved demand for private schooling. For example, they control for variation across neighborhoods in income, parental education and the taste for private schooling.

Critically, this empirical strategy requires that private school prices vary within a neighborhood. In the absence of price variation within neighborhoods, the price coefficient Equation in (4.3) is not identified. As we next describe, we have identified variation in tuition prices that occurs within neighborhoods. After describing those data, we return to defining out our empirical strategy, showing how we will use these data to identify the price elasticity of demand for private schooling.

4.4.1 Sibling Discounts at Catholic Schools

Table 1 shows the tuition prices charged by two private, Catholic schools in Columbus, Ohio. These elementary schools are quite similar in size and both enroll children in kindergarten through eighth grade. Families enrolling one child in these two schools face similar costs: St. Catherines charges \$1,125 and Blessed Sacrament \$1,200. But families seeking to enroll two children face very different costs: Blessed Sacrament charges \$1,200 for the second child but St. Catherines charges only \$325. The costs diverge still further if a family has three children it wishes to enroll: St. Catherines charges no tuition for the third child while Blessed Sacrament continues to charge its flat rate of $1,200. These sibling discounts are school-specific, applying only if siblings are enrolled in the same Catholic school in the same year.

We knew of these discounts because the first author attended a Catholic elementary school that discounted tuition for siblings. Intrigued by the possibility of exploiting this source of variation in tuition prices, we searched for a dataset or publication that documented them in detail but found no such resource. We did learn from the National Catholic Education Association (NCEA) that dioceses, the sub-national administrative unit of the Catholic Church in the US, do collect such data from the schools in their region. We contacted a few large dioceses and (after hand-entering the data) confirmed that the discounts were widespread and variable, both across and within dioceses.

We therefore broadened our data collection efforts. We contact all 168 dioceses by mail, phone, and email (see Data Appendix for details). After repeated contacts, 136 dioceses representing 90 percent of Catholic school enrollment in the US responded. Sixty dioceses agreed to participate and sent data on 1,760 schools representing 37 percent of national Catholic school enrollment. An additional 31 dioceses (24 percent of national enrollment) agreed to participate but despite repeated confirmation of their intent did not send data. The 45 dioceses that declined to participate (29 percent of Catholic school enrollment) overwhelmingly cited lack of data or staffing constraints as the reason.

After entering and examining the data, we quickly determined our analysis would focus on elementary schools. Multiple siblings can spend more time together in an elementary school (spanning eight grades) than in a high school (spanning four grades) or a middle school (spanning three to four grades). Perhaps for this reason, sibling discounts are more prevalent at elementary schools than at high schools.[3]

[3] In particular, we focus on schools that span (at least) grades one through eight, the most

Our sample of elementary schools quite closely resembles the universe of Catholic schools. Figure 2 shows the distribution of Catholic dioceses across the United States. Darkly-shaded circles depict Catholic school enrollment levels in the dioceses for which we have data.[4] Our sample tilts toward large dioceses, both because we pursued their participation most aggressively and because the smaller dioceses frequently did not have the personnel and record keeping to allow them to respond to our data request.[5] Twenty-nine percent of all Catholic schools, and 30 percent of our sample schools, are located in the Northeast. Sixteen percent of all Catholic schools, and 19 percent of our sample schools, are in the South. Forty-four percent of all Catholic schools, and 37 percent of our sample, are in the Midwest. The average Catholic elementary school enrolls 286 students in grades kindergarten through eight, while our sample schools average 296 students.

Table 2 shows tuition data (weighted by school enrollment) for elementary schools in our sample for the 1999-2000 academic year. For the first sibling, the mean tuition charged is $1,975; in all Catholic schools, the average tuition price is $2,178 (figure is for 2000-2001 and is taken from Kealey, 2002). Sibling discounts are widespread and variable. For the second sibling the mean tuition charged is $1,473 and for the third and fourth siblings the means are $1,258 and $1,103, respectively. Tuition rates for higher-order siblings are more variable than those for the first: the standard deviation in tuition is $700 for the first sibling but $743, $1,258 and $899 for the second, third and fourth siblings, respectively. Fifty-two percent of schools offer a discount for the second child, and 69 percent offer them for the third and fourth child (Table 3). Thirty-five percent of schools offer a discount of more than

common structure for Catholic elementary schools.

[4] Data from our sample and National Catholic Education Association website.

[5] In several cases research assistants traveled to a diocese to enter data from paper records into a laptop. This was worthwhile only for large dioceses.

25 percent for the second sibling, while 14 percent more than halve tuition for the second child. For third children the discounts are steeper, with tuition cut by more than half by 40 percent of schools. A quarter of schools discount tuition for the third sibling by more than 75 percent.

This table confirms that the sibling discounts are widespread and variable.[6] Both statistical properties are critical for our identification strategy. Were sibling discounts rare, we would be unlikely to pick up their effects in the household survey data that we use to measure private school attendance. Were sibling discounts uniform, we would have difficulty disentangling their effect on private school attendance from any (perhaps nonlinear) relationship between family size and private school attendance. The spatial variation in sibling discounts will allow us to control flexibly for family structure while still identifying the relationship between price and private school attendance.

4.4.2 Exogeneity of Multiple-Child Discounts: Qualitative Interviews

The typical Catholic elementary school is affiliated with a local parish that subsidizes the schools operation. Parishes that subsidize their schools more heavily charge lower tuition prices. Until the mid-1960s, these subsidies were close to 100 percent and the typical Catholic elementary school charged no tuition. At of 2001, parish subsidies covered just 24 percent of per-pupil expenditures for the 85 percent of schools that receive them (Kealey, 2002). Parishioners that heavily subsidize

[6] Most of this variation occurs within dioceses. A regression of tuition charged for the first child against a set of diocesan fixed effects yield an R^2 of 0.35, indicating that just 35 percent of the variation in tuition prices is explained in differences across dioceses in their average tuition rates. Sixty-five percent of the variation is therefore within dioceses, indicating that the schools have substantial autonomy in setting their prices.

schools may be parishioners that especially desire Catholic schooling. It is exactly this endogenous price-setting that motivates our search for alternative identification strategies for identifying the price elasticity of demand for Catholic schools.

In interviews, we asked Catholic school principals how they set prices. They typically replied that they assess their costs and parish subsidy and then choose a (first-child) tuition price that will allow them to break even. Administrators never volunteered how they set sibling discounts. In response to our probes about sibling discounts, the rationale most frequently offered was that schools hope to be affordable to large families. By offering a "family rate," a school might convince a household to send all of its children to that school.[7] Several respondents volunteered there were relatively few sibling pairs or triplets in their school, so that even large discounts did not have much impact on overall revenue. In summary, it appears that the setting of first-child price is treated as a financial decision while prices for subsequent siblings are perceived as a service to local families.

4.4.3 Estimating Equation

We can now describe how we will use sibling discounts in our estimation strategy. Consider families of varying sizes that live near either St. Catherines or Blessed Sacrament. Families enrolling one child in private school face similar tuition costs in these two neighborhoods (Table 4, $1,125 versus $1,200). By contrast, families enrolling two children face very different costs ($1,450 near St. Catherines and $2,400 near Blessed Sacrament). The difference in costs between families near Blessed Sacrament and St. Catherines is $75 for families enrolling one child and $950 for

[7] Personal communications with Sister Mary Taymans of the National Catholic Educational Association, September 11, 2002 and Sister Judy Cauley of Archdiocese of Chicago, October 7, 2002.

families enrolling two children. Another way to look at the data in this table is that the difference in costs between families enrolling one vs. two children is $1,200 in the neighborhood near Blessed Sacrament and $325 near St. Catherines. The difference-in-difference of total tuition costs is $875.This difference-in-difference in tuition costs forms our identifying source of variation in price.

In practice, we execute this strategy by controlling for neighborhood fixed effects and family composition fixed effects:

$$y_{jnb} = \beta_0 + \beta_1 \text{price}_{nb} + \delta_n + \delta_b + \varepsilon_{jnb} \tag{4.4}$$

In this equation, y_{jnb} indicates the private school choice of family j, with composition n, that lives in neighborhood b. δ_b indicates a set of neighborhood fixed effects. We experiment with a variety of neighborhood definitions, ranging from the census tract to the census block. We ultimately settle on census block groups as our definition of a neighborhood. We describe census block groups in the next section. With the block-group fixed effects, we non-parametrically control for any unobserved differences in the demand for private education across block groups, such as the poverty rate, crime rate and population density. If schools respond to local preferences in choosing the level of their prices, then our fixed effects strategy will eliminate bias in the estimated demand elasticity.[8]

In order for our price coefficient to be identified from the interaction of family composition and local tuition schedules, we must include flexible controls for the main effect of family composition. The discounts a family can obtain are a function

[8] Note that this approach controls for any fixed quality differences across schools as well so long as there is not more than one school per block group. Since some block groups do have more than one closest school, we have also estimated our models with school fixed effects; the results are slightly larger but not significantly different.

of the number of children that are simultaneously of elementary-school age. A family with two children spaced eight years apart would qualify for no discount, since the children would never be in elementary school at the same time. A family with two children spaced two years apart would get the second child discount for the six years that the siblings elementary school attendance overlapped. As this example makes clear, the spacing of children, as well as their number, affects the size of the familys tuition discount.

In Equation (4.4), δ_n denotes a vector of dummy variables measuring the age, number and spacing of children in a family. The dummies are constructed as follows. We calculate the age span between each adjacent sibling. For example, in a family with children of ages 3, 6 and ten, the age spans are three and four years. We then define a set of dummies that define the number of age spans of a given width in each family, and include these in the regression.[9] We also include a set of eighteen age dummies that indicate the presence of children age 0, of age 1 ... of age 18 in the household. These variables for the number and spacing of children will eliminate from the identifying variation in price the average, nationwide sibling tuition discounts. They also control for any nationwide correlation between family composition and private school attendance.

Our key explanatory variable of interest is price$_n b$, the total cost to a family of composition n in neighborhood b of sending all of its children to the nearest Catholic elementary school. This price is a function of the number and spacing of children in a family as well as the neighborhood in which the family resides. Our key outcome of interest is an indicator for whether all of the elementary-school-age children in family j are enrolled in private school. We have chosen to define cost, and private

[9] There are 27 spacing dummies in the equations, indicating up to three occurrences of nine different age spans (0 to 8+).

school attendance, at the level of the family for two reasons. First, the schools define prices at the level of the family, rather than the individual child. Second, the data indicate that families make schooling choices at the level of the family, rather than the individual child: the overwhelmingly majority of families send either all or none of their children to private school. In the 2000 Census, among families with children of elementary school age that send any child to private school, 86 percent send all of their children to private school.

The identifying variation in schooling costs in Equation (4.4) has an intuitive interpretation: the equation is identified by within-neighborhood differences in the total cost of sending a familys children to the local Catholic school. Differences in total costs have a natural economic interpretation as marginal costs. The thought experiment is that (within a neighborhood) families are randomly assigned a sibling age structure, which generates variation in the total cost of sending a familys children to private school. In this thought experiment, the marginal cost of private schooling within a neighborhood is the cost of being assigned one sibling structure vs. another sibling structure. And since our identification comes from the interaction of family structure with tuition schedules, an equivalent thought experiment is that families with a given sibling structure are randomly assigned a sibling discount schedule. In this thought experiment, the "marginal cost" of private schooling varies across neighborhoods, and is the cost of being assigned one tuition schedule vs. another.

4.4.4 Data on Private School Attendance: Restricted Census of Population and Housing

Our estimation strategy requires data on childrens private school attendance, as well fine geographic identifiers that allow us to link a household to the nearest private school. The Public Use Microdata Sample (PUMS) files of the 2000 Census of Population and Housing collects data on school enrollment for all household members who are age three and above. These enrollment variables capture whether i) an individual has attended school in the past two months and ii) whether that school is public or private. In Census 2000, 7.8 percent of families have all of their elementary-school age children enrolled in private school.

We conduct our analyses at the level of the Census sub-family. We construct measures of the number and ages of the children in each subfamily (hereafter referred to as a family), as well as the education, race and ethnicity of the mother and father, if present. Our analytic sample is restricted to families that contain no more than six children below age 19 and no more than three children between the ages of six and thirteen. This restriction excludes only two percent of families with any children between age six and thirteen.[10]

For reasons of confidentiality, fine geographic identifiers are not contained in the public-use versions of the Census. We analyze restricted-use versions of the PUMS that contain geographic identifiers at the level of the block. A census block is the finest geographic unit used by the Census Bureau and is its closest approximation to a neighborhood. There are about 8 million blocks in the US, ranging in population from zero to a few hundred (three million blocks are empty). Using block identifiers, we matched each family to its closest Catholic elementary school. Distance was calculated using mapping software, as the crow flies, from the population-weighted centroid of the block to each schools exact address.[11] If the closest Catholic elementary school to a block was not in our analytic tuition sample, we discarded the

[10] We also suspect that some of the largest "families" are actually misclassified group quarters.

[11] The Census Bureau maintains a dataset of the latitude and longitude of each block centroid. We calculated the latitude and longitude (of the physical location) of each Catholic school using mapping software.

block. A Catholic elementary school would not be in our sample for one of three reasons: the block is located in a diocese that did not give us data, the school is not administered through the Catholic diocese, or the school has a non-traditional grade structure (e.g., grades K through 5, grades 6 through 9).[12] More details of the mapping and matching process are in the Data Appendix.

In our preferred specification, we control for block-group fixed effects. There are 213,607 block groups in the continental United States. A block group is a subdivision of a Census tract. Block groups typically contain 1,500 people, with a Census-defined minimum of 600 and maximum of 3,000. The typical person in the US lives in a county that contains over 700 block groups.[13] Block groups are intended to be spatially-coherent units, whose boundaries consist of "visible and identifiable features, such as roads, rivers, canals, railroads, and above-ground high-tension power lines."[14]

In the 1-in-6 sample of the PUMS, sub-families that meet our sample restrictions concerning the ages and number of children reside in 1,736,984 blocks that are contained in 206,703 block groups (Table 5). These 2,969,515 families include 4,235,364 children of elementary school age. About sixteen percent of these families (463,505) live within ten miles of a Catholic elementary school for which we have tuition data for the years 1999, 2000 or 2001. These 463,505 families form our analytic sample.

As can be seen in Table 5, our analytic sample (Column 2) is fairly similar to the broader sample of block groups contains children of elementary school age. Unsurprisingly, private school attendance is higher in our sample (13.4 percent) than

[12] For the dioceses that sent us data, we very rarely lack tuition data for any of the schools in its catchment area.

[13] This is the (population-weighted) average number of block groups in a county in 2000.

[14] This paragraph's description of Census geographic areas is taken from US Bureau of the Census, "Census 2000 Statistical Areas Boundary Criteria," http://www.census.gov/geo/www/psapage.html#BG, accessed January 26, 2007.

in the full sample (7.8 percent). Family size and parental race and education are similar in our analytic sample and the full Census sample of households with children between six and thirteen. Mean income is slightly higher in our analytic sample, likely reflecting the fact that our sample tilts toward urban areas and the Northeast, and away from suburbs and the South. This reflects the spatial concentration of Catholic schools.

4.4.5 Measurement Error in Catholic School Attendance

The census school enrollment variable does not specify whether that private school is Catholic; that question was last fielded in the 1980 Census of Population and Housing.[15] How does this affect the interpretation of the price coefficients in our estimating equations? It is helpful to write private school attendance as the sum of Catholic school attendance and non-Catholic private school attendance:

$$y = y^{\text{cath}} + y^{\text{non-cath}}$$

Subscripts are suppressed to simplify exposition. This identity holds for families as well as in the aggregate. Plugging this identity into our key estimating equation and rearranging terms yields:

$$\begin{aligned} y^{\text{cath}} + y^{\text{non-cath}} &= \beta_0 + \beta_1 \text{price} + \varepsilon \\ y^{\text{cath}} &= \beta_0 + \beta_1 \text{price} + (\varepsilon - y^{\text{non-cath}}) \end{aligned}$$

We see that $y^{\text{non-cath}}$ is contained in the error term. If $y^{\text{non-cath}}$ is uncorrelated

[15] From the US Department of Educations Private School Survey, we do know that about half of private school attendance is in Catholic schools.

with price (conditional on neighborhood and family structure fixed effects) then mismeasurement in Catholic school attendance will not bias $\hat{\beta}_1$ so that $\hat{\beta}_1$ is an unbiased estimate of the relationship between Catholic school prices and Catholic school attendance. This condition holds if non-Catholic schools do not offer sibling discounts, or if they offer discounts that are uncorrelated with those offered by Catholic schools. This condition also holds if non-Catholic private schools offer sibling discounts that are uniform at the national or regional level (e.g., if the schools use the need-based financial aid formula promulgated by the Private School Scholarship Service, which incorporates a discount for larger families). Any such uniform discounts would be absorbed by our family composition fixed effects. At the opposite extreme, if non-Catholic schools offer sibling discounts identical to those offered by nearby Catholic schools, then $\hat{\beta}_1$ is an unbiased estimate of the relationship between private school prices and private school attendance, since Catholic school prices act as a perfect proxy for the tuition charged by non-Catholic private schools.

Our results suggest that the results are driven by Catholic school attendance and Catholic school prices. As we show later in the paper, the price effects are much larger among those who (based on ethnicity) are most likely to be Catholic.

4.5 Results

The baseline results are in Table 6. We start with a bivariate regression that includes on the right-hand side only the family cost variable. We have multiplied the price coefficient by 100 to allow for ease of interpretation. The coefficient of -0.072 in Column (1) indicates that an increase in tuition cost of \$1000 is associated with about a tenth of a percentage point decrease in the probability of private school attendance. This coefficient is neither substantively not statistically different from zero. The equation has very little explanatory power, with a R^2 of essentially zero.

This specification does not control in any way for family composition. Families with more children face higher total costs, and they may be more (less) likely to send their children to Catholic school. This would tend to produce a positive (negative) bias on the estimated coefficient. We therefore add to this bivariate regression variables (described in the previous section) that capture the ages, number and spacing of a familys children. These variables net out differences in price and private school attendance across children of different ages and families of different compositions. The coefficient of 0.454 in Column (2) indicates that, conditional on family composition, an increase in a familys tuition costs of \$1000 is associated with a 0.454 percentage point increase in the probability of the family sending all of its children to private school. The coefficient is highly significant, with a standard error of 0.08 percentage points.

We next add to the specification plausible, observable determinants of demand: income, parents education, ethnicity, race and parents' marital status.[16] This set of covariates has some explanatory power: the R^2 rises from to .05 when they are added to the regression. With the addition of these covariates the price coefficient is once again small, negative and insignificant: -0.049 with a standard error of 0.075. This is small both statistically and substantively.

The price coefficient in these specifications is identified, in part, by variation across neighborhoods in the price of the nearest school. The zero-to-positive price coefficient likely reflects the bias predicted by a simple model of supply and demand: across neighborhoods, equilibrium levels of tuition prices and enrollment are deter-

[16] Demographics consist of dummies for: mother's and father's education (less than high school, high school, some college, college grad); presence of mother and father; mother's and father's marital status; mother's and father's race and ethnicity; and family income (\$10K brackets, with \$200K+ a single bracket).

mined both by local demand shocks, which move us along a positively-sloped supply curve, and local supply shocks, which move us along a negatively-sloped demand curve.

Geographic fixed effects allow us to control for any unobserved (and observed) determinants of demand that vary across neighborhoods. If families with similar tastes for private schooling live near each other, these fixed effects will have substantial explanatory power in our regressions. Note that our use of neighborhood fixed effects is feasible only because multiple-child discounts create variation in tuition costs within neighborhoods. The use of such fixed effects has not been possible in previous research, in which tuition costs have varied only across state or school district.

We start with a set of tract fixed effects; there are 16,609 tracts in our data. Since tract population varies from 1,500 to 8,000, this is a very loose definition of a neighborhood. But even this crude measure of geography explains more than twice as much of the variation in private schooling as observable characteristics: the R^2 in a regression that includes tract effects but no demographics is 0.133 (Column 4), while that for the regression including demographics but no tract effects is 0.05 (Column 3). More importantly, the tuition coefficient becomes substantially more negative and is now highly significant (-0.235, with a standard error of 0.104). With the addition of covariates to this tract-effects specification, the coefficient is increases in magnitude (to -0.295) and is slightly more precise (standard error of 0.100). This increase indicates that, even within a tract, observable family attributes are correlated with both price and school attendance. Our data allow us to include block-group fixed effects, an even finer level of geography than tract. The typical census tract contains three census block groups; there are 42,226 block groups in our sample. The R^2 in a regression that includes block-group effects but not demographics is 0.207 (Column

6), as compared to 0.133 for the tract-effects specification. The magnitude of the price coefficient increases to -0.356 when block-group fixed effects are included. This coefficient indicates that a $1,000 increase in a familys tuition costs decreases the probability that its children attend private school by 0.36 percentage points. The coefficient is precisely estimated, with a standard error of 0.12 percentage points. Once block group fixed effects are included, the cost coefficient is insensitive to the inclusion of demographic variables: the coefficient is -0.38 with their inclusion (Column 7) and -0.36 when they are excluded (Column 6).

4.5.1 Are Families Myopic or Forward-Looking in Their Schooling Decisions?

We next explore alternative specifications of the price variable. We have so far assumed that families are essentially myopic, considering only current tuition costs when deciding whether to enroll their children in private school. These present costs incorporate sibling discounts for children that are currently of elementary school age, but they ignore any discounts that are produced by the private school attendance of siblings who are currently older or younger than elementary school age (that is, under six or over thirteen). A forward-looking family would consider not only todays tuition costs, but the lifetime costs of private school, which would incorporate discounts generated by all siblings in the family. In this section we show results based on these two models of family decision-making and statistically test which model better fits the data.

Consider a family with m children of which n_t are of elementary school age at time t.[17] For example, assume a family that, on Census day, has three children

[17] Census measures the number of children in a family with error, since the youngest may not

aged, 3, 6, and 10. Their closest Catholic school charges $2,000 for the first enrolled sibling, $1,500 for the second and $1,000 for the third. We define a myopic family as one that decides whether its children will attend private school this year based on the current costs of sending n_t children to Catholic school this year. This cost incorporates multiple-child discounts, but only for the n_t children of elementary-school age. In our example, the middle and oldest child are of elementary school age but the youngest is not, so the myopic cost in 2000 is $3,500, the price charged a family with two simultaneously-enrolled siblings.

We define a forward-looking family as one that decides at the time of the school entry of its first-born child whether to send all of its m children to private school from grades one through eight. In this model, the salient cost is that of sending m children to private school for eight years. This cost incorporates multiple-child discounts for all m children in the family, whether or not they are currently of elementary-school age.[18] In the forward-looking model, the salient cost for our sample family is $39,000. A forward-looking family may weigh future costs less heavily that present costs (a myopic family is a limiting case, giving future costs a weight of zero). If our sample family discounts the future at a rate of 3 percent a year, their lifetime, discounted tuition cost is $33,096.[19]

In the first column of Table 7, we reproduce results from the previous section, now labeling them as "myopic." In the next two columns we show results for the forward-looking model with discount rates of three percent and zero percent. In all

yet be born and the oldest may have formed their own households. If the degree of error is random across block groups, our estimates are biased downward.

[18] Future tuition schedules are unknown, of course, and we have past tuition schedules for only a subset of our schools. We therefore assume stability of tuition prices. That is, we assume that the familys best forecast of future tuition prices (and our best guess at past prices) is current prices.

[19] Here we treat Census 2000 as $t = 0$, discounting any costs going forward and inflating costs going backward.

of these specifications, the outcome of interest is the same: whether all of the children who are currently of elementary school age are attending private school. All of the specifications include block-group fixed effects and controls for the ages, spacing and number of children in the family. In the myopic model, the price coefficient in is -0.356, with a standard error of 0.122. This coefficient suggests that a $1,000 increase in current tuition costs (about a third of the average) decreases the probability of private school attendance by 0.36 percentage points. The implied elasticity of catholic school attendance (assuming uncorrelated non-catholic private school prices, as discussed above) is -0.15.

In the analogous forward-looking model, with the future discounted at an annual rate of three percent, the price coefficient is -0.44, with a standard error of 0.123. The latter coefficient suggests that a $10,000 increase in the present-discounted value of lifetime tuition costs (also about a third of the average) decreases the probability of private school attendance by 0.44 percentage points. The implied elasticity here is about 30 percent larger, at -0.19, although it is not significantly different. The model that incorporates no discount rate produces very similar results (-0.40 percentage points). Adding demographics changes none of these results substantially.

Note that present costs are nested within lifetime costs: lifetime costs are the sum of present costs, past costs and future costs. This allows us to test the myopic against the forward-looking models in a straightforward fashion. We execute regressions with two price terms: one for present costs and a second that captures past and future costs. We then test the hypothesis that the coefficient on the second term differs from zero. This test rejects the myopic model; the t-statistic on the sum of present and future costs is significant. The coefficient on the present costs is larger (-0.196) and less precise than that on past and future costs (-0.029). The results suggest that families are indeed sensitive to lifetime costs when making their schooling decisions. We will therefore focus in the rest of the paper on the forward-looking model. Since the undiscounted and discounted forward-looking models produce similar results, we focus on the undiscounted results.

4.5.2 Exogeneity of Multiple-Child Discounts

Our approach assumes that sibling discounts are set exogenously to neighborhood preferences for private schooling. There are two mechanisms that would violate this assumption. First, schools may set their sibling prices according to perceived differences in demand between smaller and larger families in the neighborhood. Second, large families with a taste for private school may choose to live near schools with large discounts.[20] Both mechanisms would generate a spatial correlation between family size and the generosity of sibling discounts.

In Table 8, we probe the data for such a correlation by testing for a relationship between the magnitude of sibling discounts and the size of nearby families.[21] In this analysis, the unit of observation is the school. The dependent variable captures sibling discounts at the school. We compactly parameterize these discounts in the following way. For each school, we calculate the cost of enrolling three children born two years apart in first through eighth grade. We then calculate a counterfactual undiscounted tuition cost for this family, by assuming that each school would charge a flat tuition rate equal to what is now its first-sibling price. We divide the discounted cost by the undiscounted cost, yielding a discounted tuition index that takes value one in a school that offers no sibling discounts. This index averages 0.85, indicating that the "typical" school discounts lifetime tuition costs by fifteen percent for our

[20] If smaller families with a taste for private schooling also live close to schools with large discounts, there is no threat of bias, since the block group fixed effects control for any preferences shared by large and small families.

[21] That is, families for which this school is the closest Catholic school.

hypothesized family. The 25th percentile is 0.74 and the minimum value is 0.45. One quarter of schools offer no discounts at all and so their value is one.

To test for a correlation between the size of discounts and family size, we regress the schools discount index against variables measuring the size of nearby families. A non-zero coefficient suggests that family size and school discounts are indeed correlated. The unit of observation in these regressions is the school; there are 1,760 in our sample. "Nearby" families are those for whom this school is the closest Catholic school, as we have defined it in the rest of the paper. The right-hand variable of interest measures the share of families with more than one child of elementary school age (all families in the sample have at least one such child); the mean of this variable is 0.36, while the 50th and 25th percentiles are 0.36 and 0.32, respectively. The first column shows the bivariate relationship between the discounted tuition index and family size. The coefficient is -0.048, with a t-statistic of less than one. We will interpret the practical magnitude of this coefficient shortly (and conclude that it is very, very small). The sign of the coefficient implies that the discounted tuition index is lower where families are larger; that is, discounts are larger where families are larger. However, the sign of the coefficient flips sign (to 0.29) when we control for the demographics of nearby families and region fixed effects, implying that discounts are smaller where discounts are larger; the coefficient is still insignificant.[22]

The overall picture from this table is that of a very small coefficient with a very small standard error—that is, a precisely estimated zero. In practical terms, these coefficients are miniscule, as the following calculation shows. Take the largest coefficient (0.29, in Column 2). Its magnitude suggests that an increase of one percentage point in the share of families in a neighborhood with more than one

[22] We collapse the demographics down to (family-weighted) school-level means in order to include them in this school-level regression.

school-aged child is associated with an increase in the discounted tuition index at the nearest Catholic school of 0.29. At the means of the data, increasing the share of nearby families with more than one child from 0.36 to 0.37[23] is associated with an increase in the discounted tuition index from 0.8500 to 0.8529 [=0.85+0.01(0.29)]. This corresponds to about a $100 increase in lifetime tuition costs for our imaginary family, from a base of about $33,000. We conclude from this analysis that there is no statistically or substantively significant relationship between family composition and the magnitude of tuition discounts.

The coefficients are substantively similar (very small and insignificant) when we use other metrics of the discounts (second child percentage discount, third child percentage discount) and other metrics of family size (share of families with two children, share of families with three children). These results indicate that there is no systematic relationship between the discounts offered by schools and the size of nearby families. This rules out the following threats to identification: 1) schools set discounts based on the size of nearby families 2) large families move near schools with large discounts 3) families have more children when they live near schools with large discounts.

4.6 Heterogeneity in Schooling Decisions by Parental Characteristics

We now examine whether price effects vary across demographic groups. A frequently-vocalized concern is that private schools will cream skim certain students from failing schools. This is, at its heart, a prediction about which students will

[23] One percentage point is a large increase, as the distribution of this family composition variable is quite compressed: mean 0.36, 50th percentile 0.36 and 25th percentile 0.32.

respond more elastically to the offer of a voucher. We are unable to examine how price sensitivity varies by characteristics not observed in Census, such as the degree of parental involvement in a childs school or a childs previous academic performance. We therefore cannot predict how cream skimming might occur along these dimensions. But we can measure how price sensitivity varies by race/ethnicity, parental education and income. These parameters will allow us to predict how a price subsidy to private school could alter the demographics of public and private schools.

We run pooled regressions in which the price coefficient is allowed to vary across groups. For example, in our income analysis, price is interacted with dummies that indicate whether a family is in the top, middle or bottom of the family income distribution. Each regression also includes main effects for these family characteristics, as well as the interaction of these main effects with the family composition fixed effects. This specification allows the relationship between family composition and private school attendance to vary across demographic groups, while constraining the blockgroup fixed effects to be the same across subgroups. Relaxing this latter restriction does not substantively alter the results, but does decrease precision. We first examine heterogeneity in price effects by parental education (Table 9, left panel). In our sample, two-thirds of families have a parent with any college education. Families in which neither parent attended college appear to be substantially more responsive to price (coefficient of -0.51, standard error of 0.13) than families in which a parent has attended college (-0.31, standard error of 0.13). Since the rate of private school attendance is quite low for low-education families (3.2 percent vs. 8.9 percent) the implied elasticity for low-education families (-0.51) is above five times that for highly-educated families (-0.11). These elasticities are statistically distinguishable at conventional levels.[24] The results indicate that vouchers would tend to increase the

[24] To calculate standard errors for the elasticities, we make the simplifying assumption that the

share of private school students who come from families with relatively low levels of parental education.

We next examine heterogeneity in price effects by parental race and ethnicity (Table 9, middle panel). We divide the population into three mutually-exclusive groups: Hispanics of any race, Black non-Hispanics and White non-Hispanics. Hispanics have a relatively low rate of Catholic school attendance; 3.6 percent send their children to Catholic school, compared to five percent for Black non-Hispanics and 8.1 percent for White non-Hispanics. Interestingly, Catholic school prices faced by Hispanics are about ten percent higher than those faced by the rest of the population.

White, non-Hispanic families are substantially more responsive to price: their coefficient is -0.39, as compared to -0.01 for Black, non-Hispanics and -0.19 for Hispanics. The latter two coefficients are not statistically distinguishable from zero. The implied elasticities are -0.16 (White non-Hispanics), 0.01 (Black non-Hispanics) and -0.20 (Hispanics). This is the one case when allowing the block-group effects to vary by subgroup has a non-trivial effect on the results: the Hispanic and Black coefficients flip sign, but remain insignificant (results not shown). The one unambiguous pattern that persists across specifications is that White non-Hispanics appear to be more price-elastic than Black non-Hispanics, though this difference is not always statistically significant.

Private school attendance increases with income. Ten percent of those in the top third of the income distribution send their children to Catholic school, as compared to seven and four percent in the middle and low-income groups, respectively (Table 9, rightmost panel). We estimate price coefficients for middle- and low-income families

means of tuition prices and private school attendance are population values rather than random variables. Under this assumption, the elasticities have the same statistical significance as the price coefficients, since the price coefficient is the only term in the price elasticity that has sampling variation.

that are statistically significant and large (-0.59 and -0.48, respectively), while for high-income families the effect is smaller and insignificant (-0.26). The price elasticities implied by these coefficients drop monotonically with income: -0.44, -0.27 and -0.09 for low-, medium- and high-income families, respectively. These results suggest that vouchers would increase the representation of low- and middle-income families at private schools.

We would expect that Catholic families are those most likely to take up the option of Catholic schooling. It is theoretically ambiguous, however, whether Catholics would be more or less sensitive to our identifying variation in price. On the one hand, Catholics may have such a strong preference for religious education that they are insensitive to price. On the other hand, Catholics may be most knowledgeable about (and therefore more responsive to) the sibling discounts. Catholics cannot be identified in the Census; the US government is legally barred from asking about religious affiliation in its surveys. Ethnicity is gathered, however, and this information can be used to predict religious affiliation.[25] We define terciles of the predicted probability of being Catholic (roughly, greater than 60 percent, 20-60 percent, and less than 20 percent).

The data support the hypothesis that Catholics are more sensitive to Catholic school prices (Table 10). Among those with the highest predicted probability of being Catholic, the price coefficient is -0.74, as compared to -0.12 and -0.08 for those with medium and low probability of being Catholic. The elasticities are -0.36, -0.04 and -0.05, respectively, with only the first distinguishable from zero. Note that these results provide support for the assumption (discussed earlier in the paper) that the

[25] We use the method of Gruber (2004, 2005) to generate for each family a predicted probability of being Catholic, using data from the General Social Survey, which does collect religious affiliation. This predicted probability is simply the share of the familys ethnic group that self-identifies as Catholic in the GSS. We limit the sample to non-Asian whites for this analysis.

variation in private school attendance and price drives identifies our parameters is variation in Catholic school attendance and Catholic school price.

As a compact way to summarize our predicted effects of vouchers on the demographic composition of private and public schools, we interact price with the predicted probability that a family will send its children to private school. We use demographics (race, ethnicity, income, parents education and marital status) to estimate a probit equation in which the outcome is dummy for a familys private school attendance. From these estimated coefficients we generated a predicted probability of private school attendance for each family. We then interacted dummies representing terciles of these predicted probabilities with the price variable in our preferred specification. We also include the tercile dummies as controls, as well as the interactions of the dummies with the family composition fixed effects.

The results (Table 10) indicate that families with the highest predicted probability of private school attendance are the least sensitive to price. The elasticity drops monotonically as the predicted probability of private school attendance drops: -0.09 for families most likely to attend private school, -0.28 for families in the middle of the predicted probability distribution, and -0.59 for families who are least likely to attend private school. These elasticities are statistically distinguishable from each other. These results suggest that a voucher program would disproportionately induce into private schools those who, along observable dimensions such as race, ethnicity, income and parental education, are dissimilar from those who currently attend private school. This is in marked contrast to the assumption made in previous studies (e.g., Figlio and Stone; Lankford and Wyckoff) that the new students that vouchers would induce into private school would look demographically similar to current private school students.

4.7 Conclusion

In the private schooling market, prices and quantities are equilibrium outcomes, the product of shifts along both the supply curve and demand curve. An exogenous source of variation in tuition prices is needed in order to estimate the price elasticity of demand for private school attendance. We exploit a unique and unexploited source of variation in tuition prices to estimate this price elasticity. The majority of Catholic elementary schools offer sibling discounts. These discounts reduce schooling costs for families that, in a given year, enroll more than one child in a single Catholic school. The discounts are set by individual schools and vary considerably.

As a result of these non-linear pricing schedules, a familys tuition costs are a function of the interaction of the number and spacing of their children with the pricing policies of the local Catholic school. We have collected data on these discounts from schools representing over half of Catholic school enrollment in the US. Within-neighborhood variation in tuition prices allows us to include in our demand equation extremely fine geographic fixed effects, thereby controlling for unobserved determinants of demand that vary across neighborhoods. Restricted-use Census data allows us to identify households at levels of geography down to the block. We also control flexibly for the number and spacing of children in each family, thereby absorbing any underlying relationship between family composition and private school attendance.

We find that a standard deviation decrease in tuition prices increases the probability that a family will send its children to private school by one half to one percentage point. This translates into an elasticity of the probability of private school attendance with respect to tuition costs of -0.19. Our average effect masks substantial heterogeneity in the response to price. Families with lower levels of parental education are about over four times as price elastic than other families. The price elasticity of private school attendance drops monotonically with income; it is -0.44 in the bottom tercile but near zero in the top tercile. Overall, it is those families who (along observable dimensions) are least like the current population of private school customers that are most sensitive to price, suggesting that vouchers would substantially alter the socioeconomic composition of private schools.

The offer of a voucher to students in a failing public school may well be a complex combination of treatments: the spotlight of public attention, intervention by higher levels of government in school governance, as well as a discount at a local private (or public) school. Our estimates capture only the last causal channel. But our results strongly suggest that a voucher program would disproportionately induce into private schools those who, along observable dimensions such as race, ethnicity, income and parental education, are dissimilar from those who currently attend private school.

4.8 Data Appendix

Tuition Data

In September of 2002, we began to contact Catholic dioceses, which are the sub-national administrative unit of the Catholic Church. A letter from the National Catholic Education Association, indicating its support for our efforts, was presented during these initial contacts. In our communications with dioceses we requested schools zip codes, grades taught (e.g. K-5, K-8, 9-12), total enrollment, enrollment of Catholic and non-Catholic students, and tuition schedules.

By December 2003, all 168 dioceses had been contacted by letter or e-mail at least three times and by phone at least twice. Ultimately, 45 dioceses declined to participate (29 percent of national enrollment), 60 agreed to participate and sent data (37 percent of national enrollment). An additional 31 agreed to participate but despite repeated reminders and confirmation of their intent have not sent data (24 percent of national enrollment). The remaining never responded to any of the written data requests or returned any of the multiple voice mail messages (10 percent of national enrollment). Those that declined to participate overwhelmingly cited staffing constraints or lack of data as the reason.

The data from the dioceses arrived in multiple formats: piles of paper, spreadsheets, and word-processing files. Research assistants (double) entered these data into a computer. Our sample tilts toward large dioceses, both because we pursued their participation most aggressively and because the smaller dioceses frequently did not have the personnel and record keeping to allow them to respond to our request without unduly burdening their staff. In several cases research assistants traveled to a diocese to enter data from paper records into a laptop when the diocese was unwilling to send us records. This was worthwhile only for large dioceses.

Merging Census with Tuition Data

We match our detailed tuition data to census blocks in the 2000 Census. Our matching process is as follows:

1. Calculate latitudes and longitudes for the physical location of all Catholic schools. We used mapping software to calculate the latitude and longitude of every Catholic school in the country (not just those in our tuition sample), drawing on the census of Catholic schools contained in the US Department of Educations Private School Survey.

2. Obtain latitudes and longitudes of population-weighted block centroids from Census Bureau.

3. Calculate distance from each block to every elementary Catholic school located in the same state This was necessary to limit the number of calculations.

4. Discard blocks for which distance to the closest Catholic elementary school is greater than ten miles.

5. Assign to each block the Catholic elementary school closest to the block centroid (as the crow flies).

6. Discard blocks for which the closest Catholic elementary school is not a K-8 school or is not in our tuition dataset.

Figure 1A

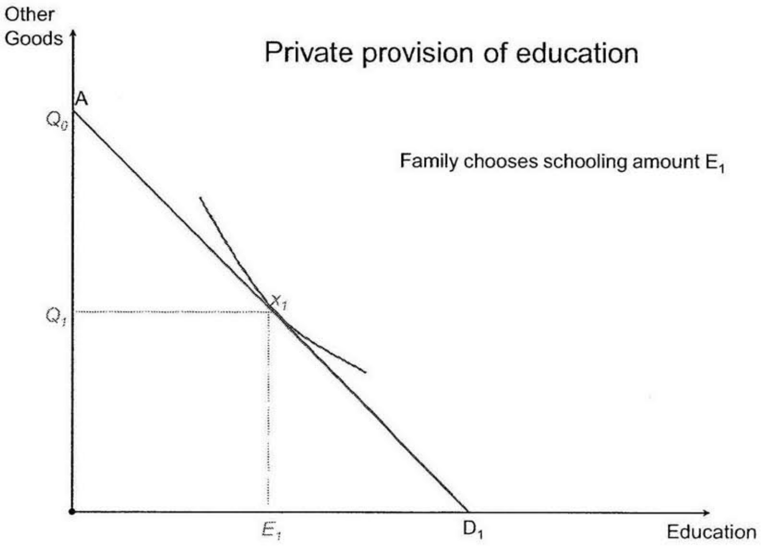
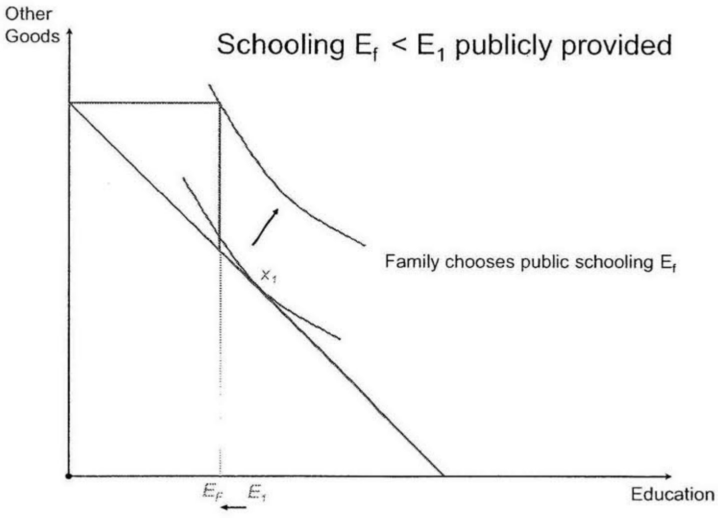

Other Goods
Private provision of education
A
Q_0
Family chooses schooling amount E_1
x_1
Q_1
E_1
D_1
Education

Figure 1B


Other Goods
Schooling E_f < E_1 publicly provided
Family chooses public schooling E_f
x_1
E_F
E_1
Education

s

Figure 1C

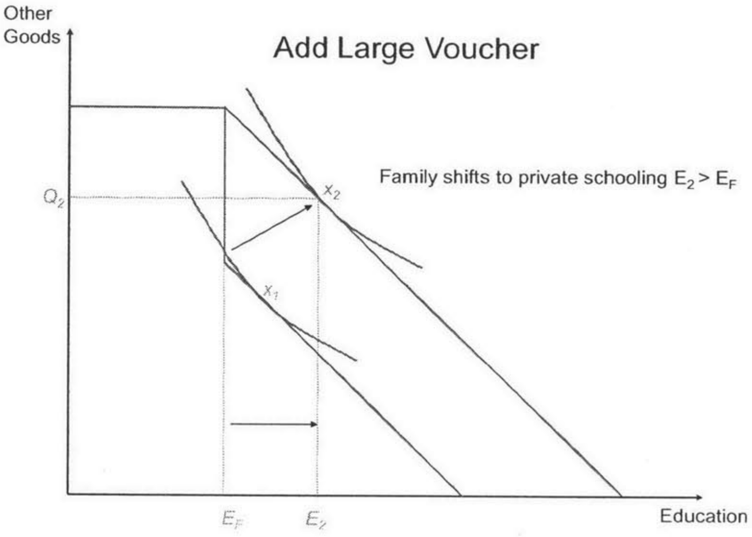


Figure 1D

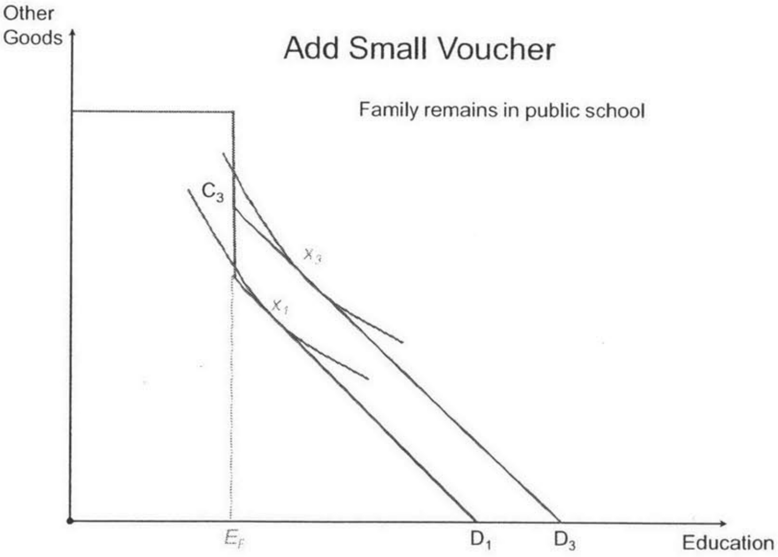


Figure 1E

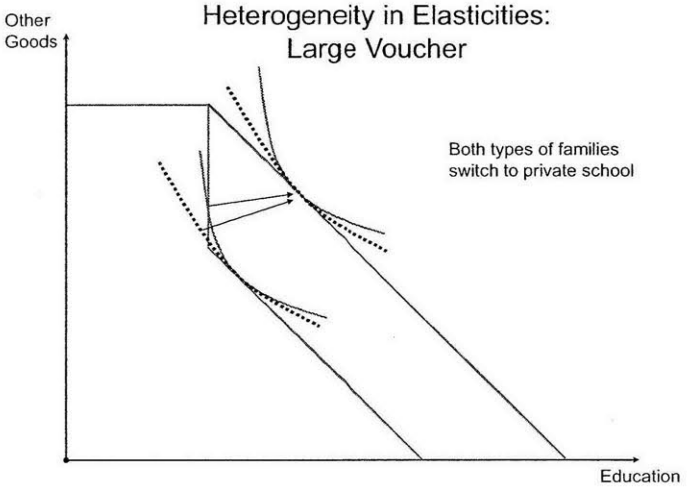

Heterogeneity in Elasticities:
Large Voucher
Other Goods
Both types of families switch to private school
Education

Figure 1F

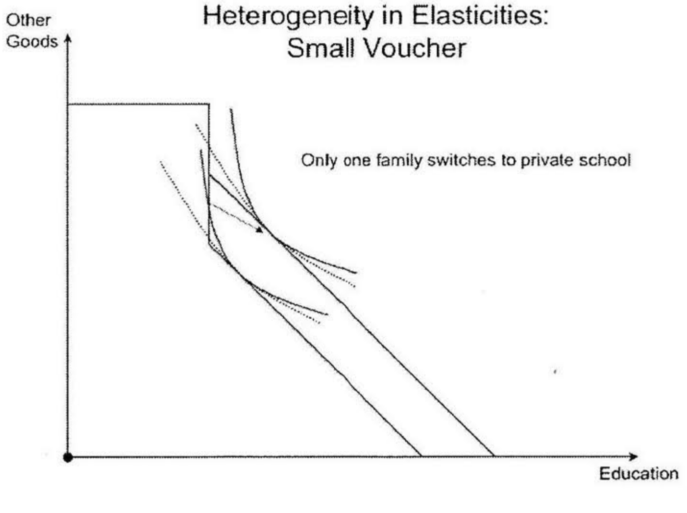
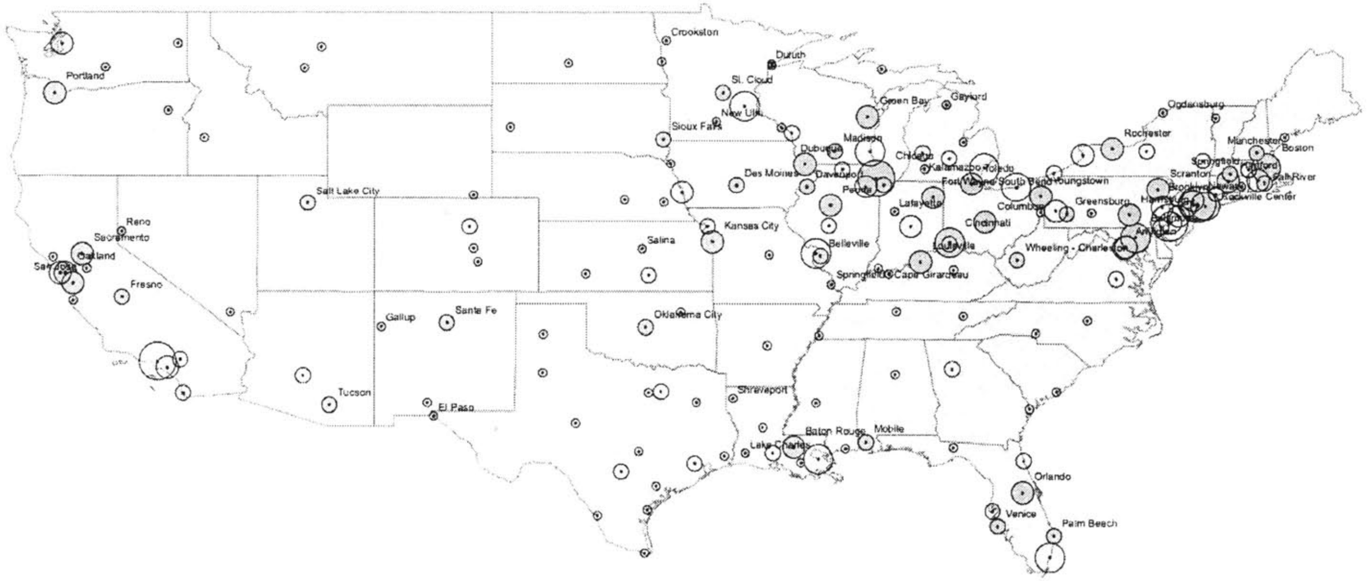

Heterogeneity in Elasticities:
Small Voucher
Other Goods
Only one family switches to private school
Education

Figure 2
Distribution of Catholic Dioceses and Tuition Sample



Note:
Green circles are dioceses in the sample, with diameter proportional to Catholic school enrollment in the diocese. Blue dots are dioceses not in sample

Table 1
Sibling Discounts at Two Schools in Columbus, Ohio

	"Blessed Sacrament"	"St. Catherine's"
Tuition, 1st Sibling	$1,200	$1,125
Tuition, 2nd Sibling	$1,200	$325
Tuition, 3rd Sibling	$1,200	0

Table 2
Catholic Elementary School Tuition Schedules
1999-2000 Academic Year
Weighted by # nearby students
N=1760

Tuition Charged for Sibling Number...	Mean	25th percentile	50th percentile	75th percentile	SD
1	1,975	1,550	1,997	2,350	700
2	1,473	965	1,400	1,860	743
3	1,258	680	1,135	1,720	1,258
4	1,103	450	1,000	1,677	899

Source: Data collected by authors.

Table 3
Shares of Schools Offering Various Sibling Discount Rates
1999-2000 Academic Year
Weighted by # nearby students
N=1,760

Discount Offered Sibling #:	None	1% to 10%	10% to 25%	25% to 50%	50% to 75%	75% to 90%	90% to 100%
2	0.48	0.06	0.1	0.21	0.11	0.02	0.01
3	0.31	0.01	0.08	0.19	0.29	0.08	0.03
4	0.31	0.01	0.06	0.13	0.22	0.08	0.17

Table 4
Difference-in-Difference in Family Tuition Costs
Two Schools in Columbus, Ohio

	One Child Enrolled	Two Children Enrolled	*Difference*
St. Catherine's	1125	1450	325
Blessed Sacrament	1200	2400	1200
Difference	75	950	**875**

Table 5 : Sample Characteristics
2000 Household Census Microdata, 1-in-6 Sample
Families with 1-3 children aged 6-13, no more than 6 children aged 0-18
Means are family-weighted

	Full Sample	Households located within 10 miles of school in Catholic school sample
	(1)	(2)
	Family Characteristics	
All children 6-13 in private school	0.078	0.134
Either Parent Black or Hispanic	0.285	0.285
Either Parent Attended College	0.623	0.678
Northeast	0.182	0.297
South	0.332	0.191
West	0.210	0.208
Midwest	0.276	0.305
Urban Area	0.571	0.836
Urban Cluster	0.106	0.056
Non Urban	0.323	0.108
Number of Children 0-18	2.54 (0.987)	2.32 (0.937)
Number of Children 6-13	1.59 (0.678)	1.68 (0.685)
Family Size	4.42 (1.288)	4.22 (1.213)
Household Income	62,536 (72,378)	72,678 (85,464)
	Sample Size	
Blocks	1,736,984	266,380
Block Groups	206,703	42,266
Families	2,969,515	463,505
Children Age 0-18	6,465,053	979,571
Children Age 6-13	4,235,364	658,832

Table 6: Baseline Analysis

Dependent variable: All children in family attend private school
Total cost: Annual cost of sending all children in family to nearest Catholic school
Unit of observation is a family (N=463,505)

	No Geographic Fixed Effects			Tract Fixed Effects [16,609]		Block Group Fixed Effects [42,226]	
FE for ages & spacing of children?		Y	Y	Y	Y	Y	Y
Demographics?			Y		Y		Y
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Total Cost ($1000)	-0.072 (0.056)	0.454 (0.080)	-0.049 (0.075)	-0.235 (0.104)	-0.295 (0.100)	-0.356 (0.122)	-0.377 (0.117)
R^2	0.000	0.003	0.05	0.133	0.17	0.207	0.235

Coefficients indicate percentage point change in probability of private school attendance associated with a $1000 increase in price. Number of geographic fixed effects in brackets. Where indicated, regressions include fixed effects for ages of children and age difference between adjacent siblings. Demographics consist of dummies for: mother's and father's education (less than high school, high school, some college, college grad); presence of mother and father; mother's and father's marital status; mother's and father's race and ethnicity; family income ($10K brackets, with $200K+ a single bracket). Heteroskedasticity-robust standard errors allow for correlation within block groups.

Table 7: Alternative Pricing Models
Dependent variable: All children in family attend private school
Unit of observation is a family (N=463,505)

	Myopic Model	*Forward-Looking Model Discount Rate = 3%*	*Forward-Looking Model Discount Rate = 0*
	(1)	(2)	(3)
Total Price	-0.366	-0.444	-0.400
	(0.126)	(0.123)	(0.126)
Implied Elasticity of Catholic School Attendance	-0.15	-0.21	-0.19
	(0.05)	(0.06)	(0.06)
Mean of Price Variable	$2,926	$33,096	$32,953
Catholic School Attendance Rate (%)	7.1	7.1	7.1
R^2	0.207	0.207	0.207
		Test Forward-Looking Model against Myopic Model	
Present Costs ($1000)		-0.152	-0.196
		(0.136)	(0.137)
Past and Future Costs ($1000)		-0.037	-0.029
		(0.014)	(0.015)
R^2		0.207	0.207

Coefficients indicate percentage point change in probability of private school attendance associated with a $1,000 increase in price (myopic model) or $10,000 increase in price (forward-looking models). The bottom panel reports coefficients from a regression including separate terms for present and all other costs. A significant coefficient on past and future costs rejects the hypothesis that families are myopic in their price-sensitivity. All specifications contain block-group fixed effects and controls for number, spacing and ages of children. Heteroskedasticity-robust standard errors allow for correlation within block groups.

Table 8: Are Sibling Discounts Correlated with Family Size?
Dependent variable: Discounted Tuition Index [=(discounted price/undiscounted price)]
Mean of Dependent Variable: 0.85
Mean of Independent Variable: 0.36

Demographics?		Y
	(1)	(2)
Share families near school with > one child	-0.048 (0.063)	0.286 (0.197)
Observations (=schools)	1760	1760

Coefficients indicate the change in the discount index at the nearest Catholic school associated with a one-percentage point increase in the share of nearby families with more than one child. Demographics consist of dummies for: mother's and father's education (less than high school, high school, some college, college grad); presence of mother and father; mother's and father's marital status; mother's and father's race and ethnicity; family income ($10K brackets, with $200K+ a single bracket).

Table 9 Heterogeneity in Price Effects, by Family Characteristics

Price Interactions

	Parents Any College	Parents No College	White non-Hispanic	Black non-Hispanic	Hispanic, Any Race	Top Income Tercile	Middle Income Tercile	Bottom Income Tercile
Price Coefficient	-0.31	-0.51	-0.39	0.01	-0.19	-0.26	-0.59	-0.48
	(0.13)	(0.13)	(0.14)	(0.18)	(0.18)	(0.17)	(0.15)	(0.14)
Implied Elasticity of Catholic School Attendance	-0.11	-0.51	-0.15	0.01	-0.20	-0.09	-0.27	-0.44
	(0.05)	(0.13)	(0.06)	(0.12)	(0.19)	(0.06)	(0.07)	(0.16)
Mean of Price Variable	$30,986	$32,282	$32,017	$33,881	$37,199	$34,731	$31,530	$32,651
Catholic School Attendance Rate (%)	8.9	3.2	8.1	5.0	3.6	10.3	7.0	3.6

Each panel is a single regression. All specifications contain block-group fixed effects, subgroup main effects, family structure effects, and the interactions of subgroup effects with family structure effects. Heteroskedasticity-robust standard errors allow for correlation within block groups.

Table 10: Heterogeneity in Price Effects, by Family Characteristics

Price Interactions

	Predicted Probability Catholic N=318,582			Predicted Probability Children Attend Private School N=440,343		
	High	Middle	Low	High	Middle	Low
Price Coefficient	-0.74	-0.12	-0.08	-0.29	-0.64	-0.50
	(0.22)	(0.19)	(0.22)	(0.16)	(0.15)	(0.13)
Implied Elasticity of Catholic School Attendance	-0.36	-0.04	-0.05	-0.09	-0.28	-0.59
	(0.10)	(0.07)	(0.10)	(0.05)	(0.07)	(0.15)
Mean of Price Variable	$34,280	$31,081	$36,158	$34,806	$30,389	$33,246
Catholic School Attendance Rate (%)	7.1	8.5	5.6	11.2	6.9	2.8

Each panel is a single regression. All specifications contain block-group fixed effects, subgroup main effects, family structure effects, and the interactions of subgroup effects with family structure effects. Heteroskedasticity-robust standard errors allow for correlation within block groups.

Bibliography

Altonji, Joseph, Todd Elder, and Christopher Taber (2005). "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling." *Journal of Human Resources*, 40(4): 791-821.

Altonji, Joseph, Todd Elder, and Christopher Taber (2005). "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy*, 113, 151184.

Altonji, Joseph, and Charles Pierret. (2001) "Employer Learning and Statistical Discrimination." *Quarterly Journal of Economics*, 116(1): 313-350.

Autor, David and David Scarborough. (2008) "Does Job Testing Harm Minority Workers? Evidence from Retail Establishments." *Quarterly Journal of Economics*, 123(1): 219-277.

Azoulay, Pierre, Toby Stuart, and Yanbo Wang. (2011) "Matthew: Fact or Fable?" Working paper. Available online: http://pazoulay.scripts.mit.edu/docs/shmatus.pdf

Ballou, Dale. (1996) "Do Public Schools Hire the Best Applicants?" *The Quarterly Journal of Economics*, 111(1): 97-133.

Ballou, Dale, and Michael Podgursky. (1995) "What Makes a Good Principal? How Teachers Assess the Performance of Principals." *Economics of Education Review*, 14(3): 243-252.

Ballou, Dale, and Michael Podgursky. (2002) "Seniority, Wages and Turnover Among Public School Teachers". *Journal of Human Resources*, 37(4): 892-912.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. (2009) "Social Connections and Incentives in the Workplace: Evidence from Personnel Data." *Econometrica*, 77(4): 1047-1094.

Becker, Gary. (1957) *The Economics of Discrimination.* Chicago: University of Chicago Press.

Bertrand, Marianne and Antoinette Schoar. (2002) "Managing With Style: The Effect Of Managers On Firm Policies." *The Quarterly Journal of Economics* 118(4): 1169-1208.

Besley, Timothy, and Stephen Machin. (2008) "Are Public Sector CEOs Different? Leadership Wages and Performance in Schools." http://econ.lse.ac.uk/staff/tbesley/papers/pubsecceo.pdf.

Billger, Sherrilyn. (2007) "Principals as Agents? Investigating Accountability in the Compensation and Performance of School Principals." *Industrial and Labor Relations Review*, 61(1): 90-107.

Blau, Francine, Janet Currie, Rachel Croson, and Donna Ginther. (2010) "Can mentoring help female assistant professors? Interim results from a randomized trial." *American Economic Review*, 100(2): 348-352.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts. (2011) "Does Management Matter? Evidence from India." NBER Working Paper #16658.

Boyd, Daniel, Hamilton Lankford, Susanna Loeb, and James Wyckoff. (2008) "The Impact of Assessment and Accountability on Teacher Recruitment and Retention: Are There Unintended Consequences?" *Public Finance Review*, 36(1): 88-111.

Branch, Gregory, Eric Hanushek, and Steven Rivkin. (2009) "Principal Turnover and Effectiveness." Paper presented at the annual meeting of the Allied Social Science Associations, Chicago.

Buddin, Richard, Joseph Cordes and Sheila Nataraj Kirby. (1998) "School Choice in California: Who Chooses Private Schools?" *Journal of Urban Economics*, 44(1): 110-134.

Calvo, Naomi. (2007) "How parents choose schools: a mixed-methods study of public school choice in Seattle." mimeo, Harvard University.

Ceci, Stephan and Wendy Williams. (2011) "Understanding current causes of women's underrepresentation in science." *PNAS*, 108(8): 3157-62.

Chandra, Amitabh and Douglas Staiger. (2010) "Identifying Provider Prejudice in Healthcare." NBER Working Paper #16382.

Chiang, Hanley. (2009) "How Accountability Pressure on Failing Schools Affects Student Achievement." *Journal of Public Economics*, 93(9-10): 1045-1057.

Chiswick, Barry and Stella Koutroumanes. (1996) "An Econometric Analysis of the Demand for Private Schooling," *Research in Labor Economics*, 15: 209-237.

Clotfelter, Charles, Helen Ladd, and Jacob Vigdor. (2006) "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *Journal of Human Resources*, 41(4): 778-820.

Committee on Science, Engineering, and Public Policy. (2011) *Expanding Underrepresented Minority Participation: America's Science and Technology Talent at the Crossroads* National Academies Press, Washington, DC.

Congressional Budget Office. (2006) "Research and Development in the Pharmaceuticals Industry." Available online at: http://www.cbo.gov/ftpdocs/76xx/doc7615/10-02-DrugR-D.pdf

Coleman, James, Thomas Hoffer and Sally Kilgore. (1982) *High school achievement: Public, Catholic and Private Schools Compared.* New York: Basic Books.

Crawford, Vincent and Joel Sobel. (1982) "Strategic Information Transmission." *Econometrica*, 50(6):1431-1451.

Cullen, Julie and Michael Mazzeo. (2007) "Implicit Performance Awards: An Empirical Analysis of the Labor Market for Public School Administrators." Working Paper.

Dee, Thomas, and Brian Jacob. (2009) "The Impact of No Child Left Behind on Student Achievement." NBER Working Paper #15531.

Downes, Thomas and Shane Greenstein (1996) "Understanding the Supply Decisions of Nonprofits: Modeling the Location of Private Schools," *RAND Journal of Economics* 27(2): 365-390.

Downes, Thomas and Shane Greenstein (2002) "Entry into the Schooling Market: How is the Behaviour of Private Suppliers Influenced by Public Sector Decisions?" *Bulletin of Economic Research* 54(4): 341-371.

Downes, Thomas and Jeffrey Zabel .(2002) "The impact of school characteristics on house prices: Chicago 1987-1991," *Journal of Urban Economics*, 52(1): 1-25.

Eberts, Randall, and Joe Stone. (1988) "Student Achievement in Public Schools: Do Principals Make a Difference?" *Economics of Education Review*, 7(3): 291-299.

Eeckhout, Jan. (2000) "On the uniqueness of stable marriage matchings." *Economic Letters*, 69: 1-8.

Ellison, Glenn. (2011) "Is Peer Review in Decline?" *Economic Inquiry*, 49(3): 635-657.

Engel, Mimi, and Brian Jacob. (2011) "New Evidence on Teacher Labor Supply." NBER Working Paper #16802.

Epple, Dennis and Richard E. Romano (1998) "Competition between Private and Public Schools, Vouchers, and Peer-Group Effects," *American Economic Review*, 88(1): 33-62.

Epple, Dennis and Richard Romano (2002) "Educational Vouchers and Cream Skimming," NBER Working Paper # 9354.

Eriksen, Homer. (1982) "Equity Targets in School Finance, Tuition Tax Credits, and the Public-Private Choice," *Journal of Education Finance*, 7(4): 436-49.

Fang, Ferric and Auturo Casadevall. (2009) "NIH peer review reform—change we need, or lipstick on a pig?" *Infection and Immunity*, 77(3): 929-932.

Figlio, David and Lawrence Getzler. (2002) "Accountability, Ability, and Disability: Gaming the System," NBER Working Paper #9307.

Figlio, David, and Cecilia Rouse. (2006) "Do Accountability and Voucher Threats Improve Poorly-Performing Schools?" *Journal of Public Economics*, 90(1-2): 239-255.

Figlio, David N. and Joe A. Stone (2001) "Can Public Policy Affect Private School Cream Skimming?" *Journal of Urban Economics*, 49(2): 240-266.

Figlio, David, and Joshua Winicki. (2005) "Food for Thought: the Effects of School Accountability Plans on School Nutrition." *Journal of Public Economics*, 89(2-3): 381-394.

Garicano, Luis and Paul Heaton. (2010) "Information Technology, Organization, and Productivity in the Public Sector: Evidence from Police Departments." *Journal of Labor Economics*, 28(1): 167-201.

Gerin, William. (2006) *Writing the NIH grant proposal: a step-by-step guide.* Thousand Oaks, CA: Sage Publications.

Ginther, Donna, Walter Schaffer, Joshua Schnell, Beth Masimore, Faye Liu, Laurel L. Haak, and Raynard Kington. (2011) "Race, Ethnicity, and NIH Research Awards." *Science*, 333(6045): 1015-19.

Glazerman, Steven. (1998) "Determinants and Consequences of Parental School Choice." mimeo, University of Chicago.

Gruber, Jonathan (2004) "Pay or Pray? The Impact of Charitable Subsidies on Religious Attendance." *Journal of Public Economics*, 88(12): 2635-2655.

Gruber, Jonathan (2005) "Religious Market Structure, Religious Participation and Outcomes: Is Religion Good for You?," *Advances in Economic Analysis and Policy* 5(1).

Hall, Bronwyn. (1994) "R&D Tax Policy During the Eighties: Success or Failure?" NBER Working Paper #4240

Hanushek, Eric, John Kain, Daniel O'Brien, and Steven Rivkin. (2005) "The Market for Teacher Quality." NBER Working Paper #11154.

Hanushek, Eric, and Margaret Raymond. (2005) "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management*, 24(2): 297- 327.

Hanushek, Eric, and Steve Rivkin. (2010) "Does Teacher Job Search Harm Disadvantaged Urban Schools?" NBER Working Paper #15816.

Hastings, Justine, Thomas Kane, and Douglas Staiger. (2005) "Parental Preferences and School Competition: Evidence from a Public School Choice Program." NBER Working Paper # 11805.

Hegde, Deepak. (2009) "Political Influence behind the Veil of Peer Review: An Analysis of Public Biomedical Research Funding in the United States" *Journal of Law and Economics*, 52(4): 665-690.

Howell, William, Patrick Wolf, David Campbell, and Paul Peterson. (2002). "School Vouchers and Academic Performance: Results from Three Randomized Field Trials." *Journal of Policy Analysis and Management*, 21(2): 191-217.

Imbens, Guido and Joshua Angrist. (1994) "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467- 475.

Ichniowski, Casey, Kathryn Shaw, and Giovanna Prennushi. (1997) "The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines." *American Economic Review*, 87(3): 291-313.

Jacob, Brian. (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.

Jacob, Brian. (2010) "Do Principals Fire the Worst Teachers?" NBER Working Paper #15715.

Jacob, Brian and Lars Lefgren. (2011) "The Impact of Research Grant Funding on Scientific Productivity." *Journal of Public Economics* 95(9-10): 1168-1177.

Jacob, Brian, and Steven Levitt. (2003) "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *The Quarterly Journal of Economics*, 118 (3): 843-877.

Kane, Thomas, Stephanie Riegg and Douglas Staiger (2006) "School Quality, Neighborhoods, and Housing Prices." *American Law & Economics Review*, 8(2): 183-212.

Kane, Thomas, Jonah Rockoff, and Douglas Staiger. (2007) "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City." *Economics of Education Review* 27(6), 615-631.

Kane, Thomas, and Doug Staiger. (2008) "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper #14607.

Keeler, Andrew and Warren Kriesel. (1994) "School Choice in Rural Georgia: An Empirical Analysis" *Journal of Agricultural and Applied Economics*, 26:2.

Kealey, Robert (1993) "Stewardship and the Catholic School Tuition Program." National Catholic Educational Association: Washington, DC.

Kealey, Robert (2002) "Balance Sheet for Catholic Elementary Schools: 2001 Income and Expenses." National Catholic Educational Association: Washington, DC.

Kerr, William. (2008) "Ethnic Scientific Communities and International Technology Diffusion." *The Review of Economics and Statistics*, 90(3): 518-537.

Knapp, Michael, Bradley Portin, Michael Copland, and Margaret L. Plecki. (2006) "Leading, Learning, and Leadership Support." Monograph. University of Washington Center for Teaching and Policy and The Wallace Foundation.

Krieg, John. (2008) "Are Students Left Behind? The Distributional Effects of No Child Left Behind." *Education Finance and Policy*, 3(2): 250-281.

Krueger, Alan and Pei Zhu. (2004) "Another Look at the New York City School Voucher Experiment." *American Behavioral Scientist*, 47(5): 658-98.

Lamont, Michele. (2010) *How Professors Think: Inside the Curious World of Academic Judgment.* Cambridge, MA: Harvard University Press.

Lankford, Hamilton, Susanna Loeb, and James Wyckoff. (2002) "Teacher Sorting and the Plight of Urban Schools: a Descriptive Analysis." *Educational Evaluation and Policy Analysis*, 24(1): 3862.

Lankford, Hamilton and James Wyckoff. (1992) "Primary and Secondary School Choice Among Public and Religious Alternatives." *Economics of Education Review*, 11(4): 317-337.

Lankford, Hamilton and James Wyckoff. (2001) "Who Would Be Left Behind by Enhanced Private School Choice?" *Journal of Urban Economics*, 50(2), 288-312.

Lavy, Victor. (2008) "Does Raising the Principals Wage Improve the Schools Outcomes? Quasi-experimental Evidence from an Unusual Policy Experiment in Israel." *Scandinavian Journal of Economics*, 110(4): 639-662.

Lazear, Edward, Kathryn Shaw and Christopher Stanton. (2011) "The Value of Bosses." Mimeo. Available online at: http://economics.uchicago.edu/pdf/lazear_101011.pdf

Lerner, Josh. (1999) "The Government as Venture Capitalist: The Long-Run Impact of the SBIR Program." *The Journal of Business*, 72(3): 285-318.

Ley, Timothy and Barton Hamilton. (2008) "The gender gap in NIH grant applications." *Science*, 322(5907): 1472-4.

Loeb, Susanna, Demetra Kalogrides, and Eileen Horng. (2010) "Principal Preferences and the Uneven Distribution of Principals Across Schools" *Education Evaluation and Policy Analysis*, 32(2): 205-229

Long, James and Eugenia Toma. (1988) "The Determinants of Private School Attendance, 1970-1980." *The Review of Economics and Statistics*, 70(2), 351-357.

Martinez, Elisabeth *et. al.*. (2007) "Falling off the academic bandwagon." *EMBO Reports*, 8: 977-81.

Mayer, Daniel, Paul Peterson, David Myers, Christina Clark Tuttle and William Howell. (2002) "School Choice in New York City After Three Years: An Evaluation of the School Choice Scholarships Program." *Mathematica Policy Research Report.*

Mobius, Markus and Tanya Rosenblat. (2006) "Why Beauty Matters." *American Economic Review*, 96(1):222-235.

Merton, Robert. (1968) "The Matthew Effect in Science" *Science* 159(3810): 5663.

National Institutes of Health. (2008) Office of Extramural Research. Peer Review Process. http://grants.nih.gov.libproxy.mit.edu/grants/peer_review_process.htm.

National Science Foundation. (2007) *Beyond bias and barriers: fulfilling the potential of women in academic science and engineering* National Academies Press, Washington, DC.

Neal, Derek. (1997) "The Effect of Catholic Secondary Schooling on Educational Achievement," *Journal of Labor Economics*, 15(1), 98-123.

Neal, Derek. (2002) "How Vouchers Could Change the Market for Education," *The Journal of Economic Perspectives*, 16(4), 25-44.

Neal, Derek, and Diane Schanzenbach. (2007) "Left Behind By Design: Proficiency Counts and Test-Based Accountability." NBER Working Paper #13293.

Oates, Wallace. (1969) "The Effects of Property Taxes and Local Public Spending on Property Values: An Empirical Study of Tax Capitalization and the Tiebout Hypothesis." *Journal of Political Economy*, 77: 957971.

Peltzman, Sam. (1973) "The Effect of Government Subsidies-in-Kind on Private Expenditures: The Case of Higher Education." *Journal of Political Economy*, 81(1), 1-27.

RAND. (2005) "Is there a gender gap in federal grant programs?" RAND Brief No. RB-9147, Santa Barbara, CA.

Reback, Randall. (2008) "Teaching to the Rating: School Accountability and the Distribution of Student Achievement." *Journal of Public Economics*, 92(5-6): 1394-1415.

Reback, Randall, Jonah Rockoff, and Heather Schwartz. (2011) "Under Pressure: Job Security, Resource Allocation, and Productivity in Schools Under NCLB." NBER Working Paper #16745.

Rockoff, Jonah. (2004) "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review Papers and Proceedings*, 94(2): 247-252.

Rockoff, Jonah, Douglas Staiger, Thomas Kane, and Eric Taylor. (2010) "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools." NBER Working Paper #16240.

Roth, Alvin, and Marilda Sotomayor. (1990) *Two-sided Matching: a Study in Game-theoretic Modeling and Analysis.* New York: Cambridge University Press.

Rothstein, Jesse. (2007) "Do Value-Added Models Add Value? Tracking, Fixed Effects, and Casual Inference." CEPS Working Paper #159.

Rouse, Cecilia. (1998) "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program." *Quarterly Journal of Economics*, 113(2), 553-602.

Rouse, Cecilia, Jane Hannaway, Dan Goldhaber, and David Figlio. (2007) "Feeling the Florida Heat? How Poorly-performing Schools Respond to Voucher and Accountability Pressure." NBER Working Paper #13681.

Sampat, Bhaven and Frank Lichtenberg. (2011) "What are the Respective Roles of the Public and Private Sectors in Pharmaceutical Innovation?" *Health Affairs*, 30(2): 332-339.

Stiglitz, Jospeh. (1974) "The Demand for Education in Public and Private School Systems." *Journal of Public Economics*, 3(4): 349-385.

US Conference of Catholic Bishops "U.S. Catholic Dioceses," retrieved on January 24, 2007 from ¡http://www.usccb.org/dioceses.htm¿.

US Conference of Catholic Bishops. "Catholic Elementary and Secondary Schools: 2004-2005," retrieved on January 30, 2007 from http://www.nccbuscc.org/education/fedasst/statistics.shtml.

US Department of Education (2002) "Paige Issues Statement On Today's Supreme Court Decision On School Choice," Press release, retrieved on January 25, 2007 from http://www.ed.gov/news/pressreleases/2002/06/06272002d.html.

US Department of Education (2004) *Digest of Education Statistics.*

US Department of Education National Center for Education Statistics. (2006) *Characteristics of Private Schools in the United States: Results From the 2003-2004 Private School Universe Survey.*

West, Martin, and Paul Peterson. (2006) "The Efficacy of Choice Threats within School Accountability Systems: Results from Legislatively Induced Experiments." *Economic Journal*, 116(510): C46-C62.

Witte, John and Christopher Thorn. (1996) "Who Chooses? Voucher and Interdistrict Choice Programs in Milwaukee." *American Journal of Education*, 104(3): 186-217.

Woolley, Anita *et. al.*. (2010) "Evidence for a collective intelligence factor in the performance of human groups" *Science*, 330(6004): 686-688.