Essays in Public Health and Early Education

by

Michael L. Anderson

B.A. Swarthmore College (1999)

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 2006

Author .....
Department of Economics
May 16, 2006

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

Certified by .....
Joshua D. Angrist
Professor of Economics
Thesis Supervisor

Accepted by .....
Peter Temin
Elisha Gray II Professor of Economics
Chairman, Departmental Committee on Graduate Studies

Essays in Public Health and Early Education

by

Michael L. Anderson

Submitted to the Department of Economics
on May 16, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

This thesis studies the long-term benefits of preschool interventions, the impact of promotions on heart disease, and the effects of light trucks on traffic fatalities.

The first chapter examines the long-term effects of preschool interventions. Several influential experiments - Abecedarian, Perry, and Early Training - have convinced many economists that preschool interventions have super-normal returns. This chapter implements a unified statistical framework to present a de novo analysis of these experiments, focusing on core issues that received little attention in previous analyses: treatment effect heterogeneity by gender and over-rejection of the null hypothesis due to multiple inference. The primary finding of this reanalysis is that girls garnered substantial short- and long-term benefits from the interventions. However, there were no significant long-term benefits for boys. These conclusions change little when allowance is made for attrition and possible violations of random assignment.

The second chapter, coauthored with Sir Michael Marmot, investigates the effect of promotions on heart disease. The positive cross-sectional relationship between socioeconomic status and health is well documented, but little evidence exists regarding the causal effect of social status on health. This chapter uses data on British civil servants from the Whitehall II study. It identifies differences in departmental promotion rates as a plausibly exogenous source of variation in promotion opportunities and exploits this variation to estimate the causal effect of promotions on heart disease. The results suggest that promotions can reduce the probability of heart disease by 3 to 13 percentage points over a 15 year period.

The third chapter analyzes the traffic safety impact of the increasing popularity of light trucks. It combines estimates from a state-level panel data set with an accident-level micro data set. The results suggest that a one percentage point increase in light truck share raises annual traffic fatalities by 0.41 percent, or 172 deaths per year. Of this increase, approximately one-quarter to one-third accrue to the light trucks' own occupants, and the remaining two-thirds to three-quarters accrue to other roadway users. Using standard value of life figures, the implied Pigovian tax is approximately 4,650 dollars per light truck sold.

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

Thesis Supervisor: Joshua D. Angrist
Title: Professor of Economics

Acknowledgments

I am eternally grateful to my advisors, David Autor, Joshua Angrist, and Jonathan Gruber, for their wisdom, insight, and encouragement during the course of this research. Without their inspiration this work would not exist. I also thank Daron Acemoglu, Dora Costa, Whitney Newey, and the participants of the MIT Public/Labor Seminar, MIT Labor Lunch, and MIT Econometrics Lunch for their helpful comments. I thank my fellow MIT students - particularly Todd Gormley, David Matsa, Sandra Chan, Rema Hanna, Patricia Cortes, Guy Michaels, and Nirupama Rao - for their advice, support, and humor. I appreciate my officemates, including Liz Ananat and Josh Fischman, for contributing to a stimulating and entertaining work environment. Alan Grant allied with me in providing indigenous Mac OS X support in a computing environment that is regrettably focused on an uninspired Windows standard.

A number of individuals helped provide data to make this research possible. I am indebted to Tarani Chandola, and other members of the Whitehall II Study Group, for their help in accessing the Whitehall II data. I am grateful to Larry Schweinhart and Zongping Xiang of the High/Scope Educational Research Foundation, Frances Campbell, Elizabeth Pungello, and Richard Addy of the FPG Child Development Institute at University of North Carolina, Chapel Hill, and Craig Ramey of the Georgetown Center for Health and Education for their generous assistance in obtaining the Perry Preschool Program and Abecedarian Project data. This research also used the Early Training Project, 1962-1979 [made accessible in 1988 as microfiche and numeric data files]. These data were collected by Susan Walton Gray, and are available through the Henry A. Murray Research Archive of the Institute for Quantitative Social Science at Harvard University, Cambridge, Massachusetts [Producer and Distributor]. In addition, funding from the National Institute on Aging, through Grant Number T32-AG00186 to the National Bureau of Economic Research, and the George Schultz Fund is gratefully acknowledged.

Many people have planted the seeds that motivated me to begin this research and provided the support that enabled me to complete it. I am grateful to the Swarthmore Economics Faculty - in particular Mark Kuperberg, Philip Jefferson, Amanda Bayer, and the late Bernie Saffran - all of whom nurtured my intellectual interest in economics and social science, and to Phil Everson, who introduced me to the joys of statistics. I appreciate my friends from Miramonte High, particularly Michael Kan and Charles Kunzman, who have always accepted me for who I am and encouraged

me in what I choose to do. I thank my friends from Swarthmore College, especially Aarti Iyer, Walter Luh, and Will Tracy, who provided such an invigorating intellectual environment. And I thank Douglas Porpora and Lynne Kotranski, who have given me so much support over the past decade.

I cannot forget my grandparents, whose impact on me cannot be overstated. I particularly appreciate Victor Anderson, who to me represents the quintessential physical scientist, and the late Lester Bergman, who affected me more than he will ever know.

I am profoundly grateful to my fiance, Joann Chuang, for understanding and accepting all of the burdens that this research has imposed, for encouraging and supporting me, even when my regressions are not working and my LaTeX is not typesetting, and for inspiring me to achieve greater things.

Most of all, I thank my parents, Mary Anderson, George Bergman, and Lim Mah Hui. Their appreciation of education and intellectual growth inspired me to get my Ph.D., and their humanity guided me into this field of research. Without them, none of this would be possible.

Contents

List of Figures

List of Tables

Chapter 1

Uncovering Gender Differences in the Effects of Early Intervention

1.1 Introduction

The education literature contains dozens of papers showing inconsistent or negligible returns to publicly funded human capital investments (Hanushek, 1996). In contrast to these studies, several randomized preschool experiments report striking increases in short-term IQ scores and long-term outcomes for treated children (Schweinhart, et al., 2005; Campbell, et al. 2002; Gray, Ramsey, and Klaus 1982). These results have been highly influential and are frequently cited as proof of efficacy for many types of early interventions (for example, Cunha, et al., 2005). They have contributed to a widespread perception that the Head Start program - one of the centerpieces of American education policy - is effective and encouraged further research on preschool programs (Currie, 2001). The experiments underlie the growing movement for universal pre-kindergarten education (Kirp, 2005). They also play an important role in the debate over the optimal pattern of human capital investments, with all parties agreeing that early education is a crucial component of human capital policy (Krueger, 2003; Carneiro and Heckman, 2003).

This paper focuses on three prominent preschool evaluations: the Abecedarian Project, the Perry Preschool Program, and the Early Training Project. Beginning as early as 1962, these programs targeted disadvantaged African-Americans in North Carolina, Michigan, and Tennessee respectively. These projects stand out from others because they implement a random assignment

research design - participants were randomly assigned to treatment (preschool) or control groups.[1] This randomization overcomes the problem of confounding that affects many observational studies.[2]

Following the initial group assignment, treated children in each experiment received several years of preschool education (intensity differed across programs). Intervention continued until treated children began regular schooling. After that point, further intervention was limited to data collection.[3] Children in both treatment and control groups received a series of standardized tests, beginning before age five and lasting through their teenage years. Researchers also conducted subject interviews and examined school and government records to collect long-term follow-up data on academic, social, and economic outcomes.

Like all experiments, notable deviations from the intended protocol occurred in each study. In two experiments, attrition materialized before preschool treatment and during the collection of follow-up data. As a result, the initial randomization in treatment status was effectively contaminated. Logistical concerns in the Perry Preschool Program also prompted the reassignment of select children between treatment and control groups, further perturbing the randomization.

In addition to the departures from experimental protocol, serious statistical inference problems affect these studies. The experimental samples are very small, ranging from approximately 60 to 120. Statistical power is therefore limited, and the results of conventional tests based on asymptotic theory may be misleading. Furthermore, the large number of measured outcomes raises concerns about multiple inference: significant coefficients may emerge simply by chance, even if there are no treatment effects.[4] All of these issues - combined with a puzzling pattern of results in which early test score gains disappear within a few years and are followed a decade later by significant effects on adult outcomes - have created serious doubts about the validity of the experiments.

[1]Two other preschool evaluations utilizing a random assignment research design exist. They are the Houston Parent-Child Development Center and the Milwaukee Project. Houston PCDC did not collect data on later life outcomes and experienced high rates of attrition. Milwaukee Project used extraordinarily small samples and suffered from a scandal involving one of its primary researchers.

[2]If parents are allowed to select whether their children receive an intervention, it is likely that the children receiving the intervention will differ in important ways from the children who do not receive it. In the context of preschool education, economists typically assume that children who attend preschool come from families that are more affluent or place a higher priority on education. Observational studies can therefore be misleading because factors other than preschool intervention may confound the results. In the context of Head Start, Currie and Thomas (1995) and Garces, Thomas, and Currie (2002) address the issue by including mother fixed effects when estimating the effect of the Head Start program on early and later life outcomes.

[3]One notable exception occurred. As discussed in Section (1.2), some treated Abecedarian children also received a schooling age treatment for several years.

[4]Both Currie and Thomas (1995) and Krueger (2003) cite these inference problems as primary reasons to doubt the validity of the experimental results.

This paper has three related objectives. First, it implements a unified statistical framework to directly address concerns about sample size and multiple testing. Second, it simultaneously examines all three studies to estimate the long-term effects of preschool separately for both males and females. Finally, it performs a detailed analysis of potential threats to validity, including attrition, violation of random assignment, and clustering.[5]

The paper is organized as follows. Section (1.2) describes the data and specific details regarding each program's experimental design. Section (1.3) sets out the statistical framework and briefly discusses possible complications.[6] Section (1.4) presents results organized by outcome stage: pre-teen, teenage, and adult. Section (1.5) summarizes the main results and discusses possible explanations for the observed causal effects. Section (1.6) conducts a thorough analysis of threats to validity, including attrition, violation of random assignment, and clustering. Section (1.7) concludes. The results demonstrate that preschool intervention has significant effects on later life outcomes for females, including academic achievement, economic outcomes, criminal behavior, drug use, and marriage. The effect on total years of education is particularly strong. However, while treatment effects are sizable for females, they are minimal or nonexistent for males - a fact relevant to the design of optimal human capital policy.

1.2 Experimental Background and Data Description

1.2.1 The Abecedarian Project

The Abecedarian Project recruited and treated four cohorts of children in the Chapel Hill, North Carolina area from 1972 to 1977. Children were randomly assigned to treated and control groups.[7] The treated children entered the program very early (mean age, 4.4 months). They attended a preschool center for eight hours per day, five days per week, 50 weeks per year until reaching schooling age. The program focused on developing cognitive, language, and social skills. Children

[5]To my knowledge, I am the first independent researcher to analyze the micro data for all three programs.

[6]The complications are addressed in detail in Section 1.6.

[7]In fact, the experiment used a slightly more complex 4-way design. Children were assigned to one of four groups: preschool treated, preschool and schooling age treated, schooling age treated, and untreated. The schooling age treatment is potentially relevant: Currie and Thomas (2000) present evidence that higher quality primary schools enhance the long-term effects of Head Start. However, in this case the schooling age treatment - which included supplemental educational activities and biweekly home visits for three years - had a negligible effect, perhaps because it was not very intensive. It is therefore ignored for the purposes of this analysis. See Campbell and Ramey (1994) for further details.

in the control group received iron fortified formula, free diapers, and supportive social services when appropriate (Campbell and Ramey, 1994). Of the three preschool projects, Abecedarian was the most intensive.

The Abecedarian dataset contains 111 children. Researchers recruited 122 subjects, but 11 families declined or could not participate. Of the remaining 111 infants, 57 were assigned to the treatment group and 54 to the control group. Data collection began immediately and has continued - with gaps - through age 21.

Researchers gathered data from three primary sources: interviews with subjects and parents, program administered tests, and school records. Children received IQ tests on an annual basis from ages two through eight, and then once at age twelve and once at age fifteen.[8] Other standardized tests were also administered, but I focus on IQ scores for comparability across programs.[9] Researchers collected information on grade retention and special education at ages twelve and fifteen from school records. Data on high school graduation, college attendance, employment status, pregnancy, and criminal behavior come from an age 21 interview. Follow-up attrition rates are low for most outcomes, ranging from three to six percent in general.

1.2.2 The Perry Preschool Program

The Perry Preschool Program recruited and treated children in Ypsilanti, Michigan from 1962 to 1967. Children were randomly assigned to treated and control groups.[10] Treated children entered the program at age three and remained in it for two years.[11] The program implemented the ideas of Jean Piaget and focused on language skills, socialization, numbers, space, and time. Classes were based around activities, and teachers used conversations to help children reflect upon what they did. Children attended the program five mornings per week from October through May. Treated children also received one 90 minute home visit per week. Untreated children were interviewed for data collection, but received no other services.[12]

[8]Instead of receiving IQ tests at ages six and seven, a single IQ test was administered at age 6.5.

[9]In the externally available Abecedarian dataset, test scores outside the 5th and 95th percentiles are truncated to the 2nd and 98th percentiles. Thus the mean IQ scores reported here differ slightly from the IQ scores in the previous Abecedarian literature.

[10]The published Perry literature claims that children were matched in pairs based on initial IQ scores. One child from each pair was assigned to treatment, and the other to control. However, there is no record of the original pairing. I therefore conduct the analysis as if there were no matching of this type. If this assumption is violated, the estimated standard errors will be more conservative than necessary.

[11]One wave entered at age four and received treatment for only one year.

[12]This description is drawn mainly from Schweinhart, et al. (2005). Please see that reference for further details.

Perry researchers recruited 128 subjects in five waves. Following random assignment within each wave, pairs of children with similar IQ scores were swapped between treatment and control groups to equalize socioeconomic status and sex ratios across the two groups. A few children with working mothers were switched from the treatment group to the control group; this issue is addressed in Section 1.6. Four children in the treatment group moved away before completing preschool, and one child in the control group died. Ultimately, the treatment group contained 58 children and the control group 65, for a total sample of 123.

Researchers gathered data from four primary sources: interviews with subjects and parents, program administered tests, school records, and criminal records. IQ tests were administered on an annual basis from entry until age ten, and once more at age fourteen. Information on special education, grade retention, and graduation status was collected from school records. Arrest records were obtained from the relevant authorities, supplemented with interview data on criminal behavior. Economic outcome data come primarily from interviews conducted at ages 19, 27, and 40. Follow-up attrition rates for most variables are generally low, ranging between zero to ten percent.

1.2.3 The Early Training Project

The Early Training Project occurred in Murfreesboro, Tennessee from 1962 to 1964. Two waves of three to four year old children were randomly assigned to treated and control groups. The treated children attended preschool for ten weeks during the summer, four hours per day. The program continued until the beginning of school, for a total of two to three summers of preschool. Children received positive reinforcement in the classes and participated in activities focusing on motivation, persistence, and postponement of gratification. Treated children also received one 90 minute home visit per week for the duration of the program.[13] Control children received no treatment beyond interviews for data collection.

The Early Training Project initially gathered data on 92 children. Four children were disqualified for various reasons, leaving 88.[14] The Early Training Project differs from the other two experiments in its construction of the control group. Specifically, the study's control group consists of two distinct subsets: a local control group and a distal control group. Of the 88 children in the study, 61 lived in the town of Murfreesboro, and 27 lived in a different Tennessee town. The

[13]Home visits continued for one year after the last summer school session.
[14]Pretreatment data on the disqualified children was retained.

61 children in Murfreesboro were assigned to the treatment group with approximately two-thirds probability and the local control group with approximately one-third probability. The 27 children in the distant town formed the distal control group.

The reliance on a distal control group was an unfortunate choice in the experimental design. The two towns were not initially comparable. For example, the distal town had a higher rate of AFDC enrollment. During the project's data collection phase, trends between the two towns diverged substantially. The local town's population grew almost 25 percent, while the distal town's fell several percent. Educational outcome data also suggest that the local and distal control groups are not interchangeable. Distal control females, for instance, display a significantly higher graduation rate than local control females. I therefore drop the distal control group from my analysis, and retain only those subjects who were truly randomly assigned. This choice results in a treatment group of 43, a control group of 18, and a total sample of 65. Since the treatment and control groups are unbalanced, statistical power is even weaker than the total sample size suggests.

Early Training Project data come from three primary sources: interviews with subjects and parents, program administered tests, and school records. IQ tests were given annually from ages four through eight, and again at ages ten and seventeen. Information on grade retention and high school enrollment comes from school records. Subject interviews provide data on post-high school education status and economic outcomes. No crime data were collected. Attrition rates for most variables are below ten percent; females in particular had virtually no attrition for many variables.

1.2.4 Summary Statistics

Table 1.1 lists means and standard deviations of key variables for all three projects. The statistics highlight the degree to which these children are disadvantaged. Average IQs in the teenage years range from 93.2 to 77.6. In comparison, an IQ score of less than 70 is one criteria that the *Diagnostic and Statistical Manual of Mental Disorders - Fourth Edition* uses to define mild mental retardation. High school dropout rates range from 30 to 40 percent. In at least one sample, a majority of subjects have a criminal record. When drawing inferences regarding the results' external validity, it is important to note that the children studied are not representative of the average American child. Nevertheless, many of their attributes are not unusual for African-American youth in urban

environments.[15]

1.3 Statistical Framework and Potential Complications

1.3.1 Statistical Framework

The random assignment process makes estimation of causal effects straightforward. The primary approach compares treated children (those that received preschool) to untreated children (those that did not) across a wide variety of outcomes. In general, this difference approximates both the effect of the treatment on the treated (ETT) and the intention to treat effect (ITT). The equivalence between ETT and ITT occurs in this case because virtually every child assigned to the preschool group attended preschool, and the programs were not open to children outside the preschool group. In the language of Angrist, Imbens, and Rubin (1996), almost every member of the sample was a "complier."[16]

To conduct inference, I compute Huber-White standard errors that are robust to heteroskedasticity. Although these standard errors are asymptotically consistent, the samples are quite small - some groups contain as few as ten individuals. The Huber-White standard errors may therefore be misleading, particularly since the underlying data is distributed non-normally in some cases.[17] To address this concern, I calculate p-values that do not rely on asymptotic theory or distributional assumptions.

Instead of a standard t-test, I implement a variant of the non-parametric permutation test (Yucesan, 1995). This procedure computes the null distribution of the test statistic and requires only three assumptions: random assignment, independence, and no treatment effect. For a given sample size N_k, I draw outcomes y_i^* from the empirical distribution of y_i without replacement.[18] I

[15]For example, Miller (1992) estimates that on any given day in 1991, 56 percent of African-American males aged 18-35 in Baltimore City were under some form of criminal justice supervision.

[16]It is conceivable that some children in the control group attended *different* preschool programs. However, this is unlikely. The families in these studies were relatively poor, so it would be difficult for most of them to afford private preschool programs. The predominant public preschool program, Head Start, did not begin until 1965, and it was initially a summer program. It therefore cannot have affected results for the Early Training Project, which ended in 1964, or the Perry Preschool Program, which had no summer session. In the latter case, the data show that fewer than 20 percent of Perry children attended Head Start, and these children were distributed fairly evenly between the treatment and control groups. The Abecedarian control children, however, may have received some Head Start schooling. It would be interesting to know whether any Abecedarian control children participated in Head Start, and how their outcomes differed from control children who did not. To my knowledge this information does not exist.

[17]Horowitz (2001) demonstrates that the performance of Huber-White standard errors can be very poor in small samples. Currie and Thomas (1995) and Krueger (2003) explicitly express concerns about the small Perry samples.

[18]Since these outcomes are drawn without replacement, and the sample size is N_k, in practice I simply use the

draw binary preschool assignments z_i^* with probability $p = 0.50$ (or $p = 0.67$ in the case of the Early Training Project) with replacement. For each sample, I calculate the t-statistic for the difference in means between treated and untreated groups. I repeat the procedure 10,000 times and compute the frequency with which the simulated t-statistics - which have expectation zero by design - exceed the observed t-statistic. If only a small fraction of the simulated t-statistics exceed the observed t-statistic, I reject the null hypothesis of no treatment effect.[19]

This test is similar to several well-known tests. If the preschool assignments z_i^* were sampled *without* replacement from the empirical distribution of z_i, this procedure would generally converge to Fisher's Exact Test for binary y_i.[20] Alternatively, if the outcomes y_i^* were drawn from the empirical distribution of y_i *with* replacement, the procedure would be analogous to bootstrapping under the assumption of no treatment effect (Simon, 1997). The procedure diverges from these two techniques because it attempts to reproduce the actual experiment as closely as possible. The procedure samples the outcomes y_i^* without replacement because the original sample is not a random sample of any larger population. It samples the preschool assignments z_i^* with replacement because the original assignments were drawn with replacement.[21]

The reported p-values are correct for tests conducted in isolation, but they do not address the issue of multiple inference. Because each study examines hundreds of outcomes, some outcomes should display significance even when no effect exists. Furthermore, the small samples ensure that significant results are necessarily of notable magnitude.

I address the issue of multiple inference in three steps. First, to minimize the degree of over-testing, I choose a specific set of primary outcomes based on a priori notions of importance. Next, I implement summary index tests in three broad areas: pre-teen, adolescent, and adult outcomes.[22] Finally, I control for multiple inference at the summary index level by computing Familywise Error Rate (FWE) adjusted p-values via the free step-down resampling method.

---

original vector of y_i observations.

[19]Formally, I reject the hypothesis that the treatment has any distributional effect. For non-binary outcomes, it is theoretically possible that rejection occurs because treatment affects dispersion without affecting the mean. This seems unlikely. Furthermore, most of the outcomes of interest are binary, and anything that affects the variance of a Bernoulli random variable necessarily affects the mean as well.

[20]The procedure differs very slightly from Fisher's Exact Test in that Fisher's test rejects for small p-values while this test rejects for large t-statistics.

[21]Using alternative tests in which all sampling is done with or without replacement does not significantly affect the results.

[22]Grouping instead by type of outcome - e.g. academic, social, economic - does not substantially alter the conclusions.

The set of primary outcomes includes grade retention, special education, high school graduation, college attendance, employment, earnings, government transfers, arrests, convictions or incarcerations, drug use, teen pregnancy, and marriage. This list appears long but represents only a small fraction of all available outcomes. Nevertheless, the total number of tested outcomes exceeds 40. I therefore implement summary index tests that pool multiple outcomes into a single test.

The summary index tests originate in the biostatistics literature (see O'Brien, 1984). They are robust to over-testing because the probability of a Type I error does not increase as additional outcomes are added to a summary index. They are also potentially more powerful than individual level tests - multiple outcomes that approach marginal significance may aggregate into a single index that attains statistical significance.[23]

To implement these tests, I demean all outcomes and convert them to effect sizes by dividing each outcome by its standard deviation. This conversion normalizes outcomes to be on a comparable scale. I also switch signs where necessary so that the positive direction always denotes a "better" outcome. I then create a new variable, $\overline{s}_{ij}$, that is the mean of the normalized, demeaned outcomes.[24] Thus $\overline{s}_{ij} = \sum_{k \in \mathbb{K}_{ij}} w_k \frac{y_{ijk} - \overline{y}_{kj}}{\sigma^y_{jk}}$, where k indexes outcomes within area j, K_{ij} is the total number of non-missing outcomes for observation i in area j, $\mathbb{K}_{ij}$ is the set of non-missing outcomes for observation i in area j, and w_k is the outcome weight (weights are normalized to sum to one). I then regress the new variable, $\overline{s}_{ij}$, on treatment status to estimate the effect of preschool on area j. Any missing outcomes are ignored when creating $\overline{s}_{ij}$. This procedure therefore uses all the available data, but it weights outcomes with fewer missing values more heavily.[25]

[23] Conceptually, it is useful to consider an underlying latent index - e.g., human capital at a given age - that is expressed through multiple measures (e.g., years of education, employment, earnings, criminal record). We want to know whether preschool affects the latent index, but there are two sources of random error to contend with. First, there is error that arises from the random assignment procedure - the latent index will not be perfectly balanced across treatment and control groups in any finite sample. Second, there is random error in each outcome measure - individuals with the same latent index value may realize different values for any given outcome. Summary index tests can reduce the second source of error by combining data from multiple outcome measures.

[24] The variable $\overline{s}_{ij}$ is also weighted by the inverse of the outcome covariance matrix; the weight on each outcome is equal to the sum of its row entries in the inverted covariance matrix. This weighting increases efficiency by ensuring that outcomes which are highly correlated with other outcomes receive less weight (or, alternatively, outcomes that represent new information receive more weight).

[25] An extreme case illustrates this point. Consider an example in which one outcome is missing data for every single observation. In that case, the outcome never enters into $\overline{s}_{ij}$ for any observation, and does not affect the estimation results. An alternative estimator, detailed in Kling and Liebman (2004), simultaneously estimates the coefficients for all outcomes in a given area using a seemingly unrelated regressions (SUR) model. The general effect is computed as the mean of the coefficients in that area, and the estimate's variance is calculated using the coefficient variance-covariance matrix from the SUR model. However, this estimator drops an observation if it is missing for any outcome (with no missing outcomes, the two procedures are equivalent). Since neither test is superior on a priori grounds, I experiment with both. They return similar results, except for a few cases in which the SUR estimator is affected by

Each summary index consolidates several individual tests into a single index. However, there are still nine summary tests per gender. I therefore calculate FWE adjusted p-values for all summary index tests. Suppose that K hypotheses, $H_1, H_2, ..., H_K$, are tested. The Familywise Error Rate (FWE) is the probability that at least one of the K hypotheses in the family is rejected.[26] For summary index tests, the family of tested hypotheses is the set of nine summary index tests performed for each gender.

To adjust for FWE, I implement the free step-down resampling method (Westfall and Young, 1993). This algorithm is more powerful than simpler FWE adjustments, such as the Bonferroni Correction, because it incorporates dependence between outcomes and sequentially removes hypotheses from the family being tested as they are rejected. An example may aid the interpretation of the adjusted p-values. Consider the smallest unadjusted general effect p-value, which occurs for teenage Perry females (Table 1.4). The unadjusted p-value is approximately 0.000. The corresponding adjusted p-value, calculated via the free step-down resampling method for the entire family of female summary tests, is $p = 0.003$. Suppose we simulate the female data 10,000 times under the null hypothesis of no treatment effect. If we compute an entire set of summary effect p-values for each simulation, then the *minimum* p-value of that set will be less than or equal to the unadjusted p-value of 0.000 approximately 0.3 percent of the time. For unadjusted p-values that are above the family's minimum p-value, the family of tests effectively decreases. A monotonicity enforcement performed at the end of the procedure ensures that larger unadjusted p-values always correspond to larger adjusted p-values. The code for this procedure is detailed in Section 1.8.

1.3.2 Complications

Several complications, analyzed in-depth in Section 1.6, threaten the validity of the results. A quick summary of the complications and their resolutions follows.

Attrition is present in all three preschool experiments. If this attrition is caused by treatment status, systematic differences unrelated to the treatment could emerge between the two groups. In these experiments, the direction of the induced bias is ambiguous. To address the attrition problem,

a large number of missing observations. I therefore report results for the summary index estimator.

[26] Note that the FWE adjustment is not the same as a joint test of the hypothesis of no effect for any outcome. If a joint test rejects, we can only conclude that at least one null hypothesis is false. If the adjusted p-value rejects, we can conclude that the specific null hypothesis being tested is false. A joint test is generally more powerful, but, when it rejects, the adjusted individual test yields more information.

I impute values for key outcomes among missing individuals and examine "worst case" scenarios. Under reasonable assumptions, these imputations do not qualitatively change the paper's central conclusions.

Another complication is violation of the original random assignment. The most serious case occurred in the Perry Preschool Program; for logistical reasons, several children with working mothers in the treatment group were switched to the control group. Perry researchers did not record the identities of these children. If children with working mothers perform differently than the average child, these swaps could induce bias. I address this issue by conditioning outcomes on initial maternal employment status. I also study an entire range of possible switches that could have occurred and examine the sensitivity of the estimates to these switches. Again, the main results are unchanged.

A final complication is the possibility of dependence between observations, or clustering. In these experiments, the possibility of classroom peer effects and the systematic assignment of siblings to identical treatment groups are reasons for concern. If the peer effects or intra-family correlations are strong, the standard errors could be too small. I address the problem by estimating the results on a dataset of class-by-year means and by dropping siblings from the sample. The clustering adjustments do not substantially affect key results.

1.4 Results

1.4.1 Pre-Teen Outcomes

Preschool significantly raises early IQ scores in all experiments. It also consistently reduces early grade retention and special education placement for females, but has limited effects on grade retention and special education for males.

Table 1.2 reports effects on pre-teen IQ scores. Like all tables in this section, it presents results for both genders. For each gender, the first column reports coefficients and standard errors, the second column reports control group means, the third column reports non-parametric p-values, and the fourth column reports sample size. The last column in each table tests for differences between female and male treatment effects.

All projects demonstrate similar effects on test scores at early ages. In each project, there is

a large and significant IQ effect for at least one gender upon completion of preschool. Females continue to display a significant IQ effect at age ten in both the Abecedarian and Early Training Projects. Males, however, experience no significant IQ effect in any project at age ten.

The similarity in early IQ effects across programs occurs despite their differing intensity levels. By age five, the Abecedarian, Perry Preschool, and Early Training programs exposed children to approximately 8,000, 850, and 700 hours of preschool education respectively.[27] Nevertheless, a treatment effect that peaks at roughly ten to fifteen IQ points emerges in all three programs during the preschool years.

The results in Table 1.3 suggest that the early IQ gains translate into better performance in primary school.[28] Female grade retention falls by 20 to 30 percentage points in all three programs, with p-values ranging from 0.08 to 0.16. Female special education placement falls significantly in the Perry program (26 percentage points, $p = 0.06$) but not in the Abecedarian or Early Training programs. Males in the Abecedarian program experience a 19 percentage point decline in grade retention ($p = 0.14$) and a 27 percentage point decline in special education placement ($p = 0.06$). However, males in the Perry and Early Training programs demonstrate *increases* in grade retention of approximately 8 to 10 percentage points and no notable decrease in special education placement.

Table 1.4 reports summary index results by outcome stage and experiment. At the pre-teen stage, preschool significantly improves outcomes for females in the Abecedarian and Perry programs, with summary effect size increases of 0.45 and 0.54 respectively. Early Training females experience a summary effect size increase of 0.38; the coefficient approaches significance. Males, in contrast, do not experience consistent gains in pre-teen outcomes. Abecedarian males realize a significant summary effect size increase of 0.42. However, Perry and Early Training males experience summary effect size increases of 0.15 and 0.14 respectively; neither result approaches significance.

Gender differences in treatment effects emerge by age ten. The female IQ effects at age ten are significantly higher than the male IQ effects in both the Perry and Early Training programs. Females also experience greater drops in grade retention than males in both the Perry and Early Training programs, and the differences approach significance. Most importantly, for every exper-

[27] Currie and Neidell (2004) present evidence that higher spending increases the effects of Head Start. Although initial differences are minimal, the Early Training Project does have the lowest number of significant long-term outcomes. However, this is partially due to the relatively small samples in the Early Training Project.

[28] For Perry Preschool, the grade retention variable may contain some information on teenage grade retention. For the Early Training Project, both the grade retention and special help variables may contain some information from teenage years. For these variables, it was not possible to isolate pre-9th grade outcomes in the data.

iment the summary female pre-teen effect is higher than the summary male pre-teen effect; the difference approaches marginal significance in the Perry Preschool Project.

Although preschool positively affects pre-teen outcomes, the implications for long-term success are unclear. A short-term IQ gain may not result in any long-term economic benefit, and decreased grade retention at an early age may not affect graduation rates a decade later. For example, Currie and Thomas (1995) and Garces, Thomas, and Currie (2002) conclude that, for African-Americans, Head Start initially boosts test scores but does not have any lasting effect on academic achievement or economic outcomes. Conversely, diminishing effects on standardized tests may mask improvements in crucial non-cognitive skills that affect earnings and achievement (Heckman and Rubinstein, 2001). The next subsections focus on long-term teenage and adult outcomes.

1.4.2 Teenage Outcomes

In the teenage years, early intervention significantly improves high school graduation, employment, and juvenile arrest rates for females. However, it has no significant effect on male outcomes.

Table 1.6 presents program effects on teenage academic outcomes, including IQ scores and high school graduation rates. By age 14, initial IQ effects dissipate in all three programs. Only one IQ coefficient is statistically significant - Abecedarian males at age 15 ($p = 0.09$) - and in no case does the estimated coefficient exceed five IQ points. However, the negligible IQ effects belie strong gains among females for several important teenage outcomes.

High school graduation effects for females are sizable. Females display increases in high school graduation rates (or decreases in drop out rates) of 23 percentage points in the Abecedarian Project, 49 percentage points in the Perry Preschool Program, and 29 percentage points in the Early Training Project. The Perry result is highly significant ($p < 0.001$). The Abecedarian and Early Training results achieve or approach marginal significance ($p = 0.09$ and $p = 0.11$ respectively).[29]

In contrast, the high school graduation effects for males are weak or negative. Graduation rates *decline* by 10 and 6 percentage points for Abecedarian and Perry males respectively. Early Training males are 10 percentage points less likely to drop out, but the effect is not statistically significant.

Table 1.7 presents results for teenage economic and social outcomes. Females display positive

[29]The relative insignificance of the Early Training results is primarily a result of the relatively small sample size. The estimated coefficient is larger than the Abecedarian coefficient, but with only ten females in the Early Training control group it is difficult to conduct accurate statistical inference.

economic effects from preschool as teenagers. In Perry Preschool, treated females have teen unemployment rates that are 31 percentage points lower than untreated females ($p = 0.03$). Treated females also receive approximately 1,600 dollars less in annual government transfers at 19 ($p = 0.04$). Early Training females are 13 percentage points more likely to have worked as teens, although the effect is not significant. Males, in comparison, derive no significant economic benefits from preschool during their teenage years. Unemployment among Perry male teens is only 2 percentage points lower. Treated male teens in the Early Training Project are 6 percentage points *less* likely to have ever worked.

The preschool programs have moderate effects on teen motherhood. Abecedarian females report teen pregnancy rates that are 21 percentage points lower; the effect approaches marginal significance ($p = 0.13$). Teen pregnancy rates for Perry females are 19 percentage points lower, but the effect is insignificant. Neither Abecedarian nor Perry males experience a significant decline in the probability of teen parenthood.

Early intervention has a significant effect on female teen criminal behavior. It reduces the probability of a juvenile record by 34 percentage points for Perry females. However, this significant result ($p = 0.01$) is not mirrored among males. Perry males demonstrate an insignificant 8 percentage point reduction in the probability of arrest before age 20.

Overall, preschool has a consistent, positive effect on female teen outcomes. Teenage summary effects increase by 0.42, 0.61, and 0.55 respectively for females in the Abecedarian, Perry, and Early Training programs (see Table 1.4). The summary effect is highly significant for Perry females ($p < 0.001$) and retains significance when adjusted for multiple testing. The summary effect is also significant for Abecedarian females ($p < 0.05$), but only marginally significant for Early Training females. However, preschool has no significant effect on male teen outcomes. Summary effects increase for males by only 0.16, 0.04, and 0.10 respectively in the Abecedarian, Perry, and Early Training programs. No male summary effect approaches statistical significance.

During the teenage years, it is clear that females benefit more than males from early intervention. The female-male difference in high school graduation effects is significant in the Abecedarian Project ($t = 1.80$) and the Perry Preschool Program ($t = 3.32$). Large female-male differences also emerge among Perry teens for effects on unemployment ($t = -1.60$), criminal behavior ($t = -1.54$), and government transfers ($t = -1.96$). At the summary index level, Perry females benefit significantly

more than Perry males ($t = 3.32$). For the other two experiments, female summary effects are at least 0.25 standard deviations higher than male summary effects, although the differences are not significant. With the exception of Abecedarian IQ test scores, every reported teen effect is more positive for females than for males.

1.4.3 Adult Outcomes

At the adult stage, preschool significantly raises college attendance rates for females and appears to improve female economic outcomes and reduce criminal behavior. The effects for males, however, are weak and inconsistent. There is evidence of a modest positive effect on male economic outcomes, but it is accompanied by evidence of a negative effect on male college attendance and a mixed effect on male criminal behavior.

Table 1.8 reports treatment effects on college attendance. Preschool appears to increase the probability of college attendance for females. Abecedarian females report college attendance rates 29 percentage points higher than their control counterparts. This result is statistically significant ($p = 0.02$). Perry female college attendance rates increase by 16 percentage points, and Early Training females are 12 percentage points more likely to obtain post-high school education, although neither effect is significant.[30]

However, preschool does not appear to increase college attendance for males. Abecedarian males display a 15 percentage point increase in college attendance rates, but the effect is insignificant. Perry males are 1 percentage point less likely to attend college, and Early Training males report dramatically lower rates of post-high school education (49 percentage points lower).[31] The negative effect for Early Training males is highly significant ($p = 0.005$).[32]

Table 1.9 reports results for adult economic outcomes. Preschool has a weak but positive effect on female economic outcomes. Abecedarian women are 10 percentage points more likely to be employed at age 21. Perry females are 26 percentage points more likely to be employed at age 27

[30]Post-high school education is defined as college, vocational school, or employer sponsored education/training. For either gender, limiting the outcome to just college attendance produces coefficients of similar magnitude and significance.

[31]In cases where there is overlap, my results are similar - but not exactly identical - to the results reported in Gray, et al. (1982). The discrepancy arises because the dataset that Dr. Gray provided to the Murray Research Center does not exactly match the description of the dataset used in Gray, et al. (1982). Dr. Gray passed away several years ago, so it is unlikely that we can ever fully resolve these minor discrepancies.

[32]The most likely reason for this negative finding is multiple testing. Two other possibilities are attrition bias and negative peer effects. A detailed examination reveals both of these explanations to be unlikely. Further discussion is available from the author upon request.

($p = 0.08$). However, this effect disappears by age 40. Perry females earn more at ages 27 and 40 than their control counterparts (approximately 2,600 and 3,500 dollars respectively), but the effects are insignificant.[33] Early Training females are less likely to receive welfare at age 21, but are also less likely to receive income from work at the same age (neither effect is significant). It is possible that for Abecedarian and Early Training women, potential employment effects at age 21 are masked by increased college attendance rates. In that sense, employment data at a later age would be preferable. However, controlling for college attendance when estimating the employment effect does not appreciably change the coefficients for either program.

For males, there is no consistent evidence that preschool interventions improve long-term economic outcomes. Abecedarian males achieve an employment rate 19 percentage points higher than their untreated counterparts, but Perry males see virtually no effect on employment at age 27. Perry males do report increases in annual earnings of approximately 2,400 and 6,200 dollars at ages 27 and 40 respectively. However, all of these effects are insignificant. Perry males at age 40 experience a positive employment effect of 20 percentage points. This effect approaches statistical significance ($p = 0.11$). Early Training males, however, are *less* likely to receive income from work at age 21.

Table 1.10 presents effects on adult social behavior. Treated females report improvements for several measures of criminal behavior. Abecedarian females are 32 percentage points less likely to use marijuana ($p < 0.01$). However, Abecedarian does not significantly reduce conviction or incarceration rates for females by age 21.[34] Perry females have 86 percent fewer lifetime arrests (a reduction of 1.95 arrests, $p = 0.01$), though they are only 15 percentage points less likely to have a

[33]For both females and males, the coefficient on monthly earnings at 27 has a much higher t-statistic than the coefficient on annual earnings at 27. This difference arises because the monthly earnings coefficient is between $\frac{1}{4}$ to $\frac{1}{6}$ the magnitude of the annual earnings coefficient, rather than the expected $\frac{1}{12}$. There is no a priori reason to believe that one measure is clearly superior to the other. However, the annual earnings measure does have a lower standard deviation than the annualized monthly earnings measure. More importantly, using annual earnings at 27 instead of monthly earnings at 27 produces an estimate that is consistent with the estimated earnings differentials at age 40 using either monthly or annual measures. The implied earnings effect using annual reported earnings at age 27 is 19 percent of the control mean, while the implied earnings effect using monthly reported earnings at age 27 is 59 percent of the control mean. The implied earnings effects using annual and monthly reported earnings at age 40 are 24 and 17 percent of the control means respectively. The reported monthly earnings at age 27 therefore appear anomalous. Nevertheless, for completeness I take an average of the two measures when calculating the summary index estimator.

[34]It is tempting to assume that Abecedarian females experience no significant reduction in non-drug related criminal behavior because their underlying arrest rate is much lower than Perry females. However, this assumption is incorrect, because the Abecedarian data measures *convictions* while the Perry data measures *arrests*. Clarke and Campbell (1998) report that 43 percent of the Abecedarian sample have criminal records at age 21. 51 percent of the Perry sample have an arrest record at age 19, so the two numbers are quite comparable, particularly since the Perry sample has a higher proportion of males. Clarke and Campbell find no effect of early intervention on criminal records.

criminal record.

Treated males, in contrast, do not show significant improvements for any reported indicator of criminal behavior. Abecedarian males are slightly less likely to be convicted by age 21 or to use marijuana. Perry males are 2 percentage points less likely to have a criminal record at age 27. Perry males have 38 percent fewer lifetime arrests at age 27, but the effect only approaches marginal significance (a reduction of 2.31 arrests per capita, $p = 0.13$). The "hard" drug usage rate is 20 percentage points *higher* for Perry males, an effect which attains statistical significance ($p = 0.07$).[35]

There is some evidence that preschool affects marriage rates.[36] At age 27, Perry females have a significantly higher marriage rate than untreated females. The 32 percentage point increase represents a 382 percent rise over the control group's base rate ($p < 0.01$).[37] Perry males, however, have the same marriage rate at 27 as their control counterparts.[38]

Overall, females benefit from early intervention as adults. In the Abecedarian and Perry Preschool programs, females display positive general effects of 0.45 and 0.36 standard deviations respectively (see Table 1.4).[39] Both results are statistically significant ($p < 0.01$ and $p = 0.02$ respectively), and the Abecedarian effect is robust to FWE adjustments. However, Early Training females demonstrate no general treatment effect as adults. This could be a result of the Early Training Project's relatively short intervention program, or it could be due to low statistical power.

Unlike females, males demonstrate little evidence of positive treatment effects as adults. Summary effects for Abecedarian and Perry males increase by 0.31 and -0.02 standard deviations respectively. The Abecedarian result approaches significance, but the Perry result does not. Early Training males experience a *decline* of 0.65 standard deviations in the adult summary index. This significant decrease ($p < 0.05$) is primarily driven by low college attendance rates.

Several female treatment effects are significantly higher than corresponding male effects, al-

[35]This detrimental effect has the highest significance level of *any* of any major later life outcome measures for Perry males.

[36]Perry is the only program to date that surveys participants late enough to collect meaningful marital statistics.

[37]Interestingly, Schweinhart, et al. (2005) show that by age 40 the treated females' marriage rate is only 6 percentage points higher than the control females' rate. Part of the increase is due to divorces in the treatment group, and part is due to marriages in the control group.

[38]Again, Schweinhart, et al. (2005) show an interesting twist for males at age 40. Treated males are more likely to be married at age 40, but the entire increase is due to a larger fraction of treated males who have divorced and married multiple times. The fraction of treated males who have only married once is actually slightly lower than the fraction of controls who have only married once. It is unclear whether this pattern should be counted as a "positive" or "negative" effect.

[39]The Perry Preschool summary index also includes monthly income.

though the effect heterogeneity is less pronounced than during the teenage years.[40] The female-male treatment effect difference is significant for drug use and marriage among Perry participants ($t = -2.07$ and $t = 2.00$) and post-high school education among Early Training participants ($t = 2.35$). The difference in female-male summary effects is also significant in the Early Training Project. For drug use and post-high school education, the significance is primarily the result of negative male treatment effects rather than positive female treatment effects. Nevertheless, it still constitutes evidence of greater benefits for females.[41]

1.5 Discussion

A clear pattern emerges from a detailed examination of preschool treatment effects by gender: females display significant long-term effects from early intervention, while males show weaker and inconsistent effects.[42] Treated females show particularly sharp increases in high school graduation and college attendance rates, but they also demonstrate significant positive effects for economic outcomes, criminal behavior, drug use, and marriage.

In contrast to females, males do not appear to derive lasting benefits from early intervention. No positive, long-term outcome achieves statistical significance for males, although one, employment at age 40 for Perry males, comes close. This aggregate performance is disappointing when considering the number of outcomes tested; even with a minimal treatment effect, positive and significant results are likely to occur several times just by chance. In fact, the only significant, long-term results for key male indicators are negative.

Figure 1-1 presents a visual summary of the female-male treatment effect heterogeneity for long-term outcomes. This figure plots t-statistics for all of the reported teenage and adult coefficients across all experiments. Each point corresponds to the t-statistic for a single outcome, and all outcomes have been recoded so that the positive direction always corresponds to a "better" outcome. The first column of points plots male t-statistics, and the second column plots female t-statistics.

[40] The effect heterogeneity is reduced primarily because of a decline in the general effect size for females.

[41] The female coefficients are centered around a higher mean, so even in the face of adverse shocks they do not become negative and significant. The male coefficients, in contrast, are centered around a lower mean, and are more likely to display negative, significant effects simply due to chance.

[42] Several researchers, most recently Heckman (2005), have noted the possibility of heterogeneous treatment effects by gender in the context of Perry Preschool. However, there has been no statistical analysis of this difference, nor would it be possible to draw any strong conclusions regarding treatment effect heterogeneity by gender from Perry Preschool alone.

It is clear upon visual inspection that the distribution of female t-statistics is centered well above the distribution of male t-statistics.

The third column of points plots a set of male t-statistics generated by randomly assigning treatment status to males. This procedure guarantees that any significant "treatment effects" visible in the column are simply due to chance. The procedure is equivalent to sampling random draws from the t-distribution, except that it preserves the inherent correlation structure between t-statistics within each experiment.[43] To construct this column, I randomly generate a set of treatment assignments and compute the corresponding t-statistics. I then plot these t-statistics.

A comparison of the first and third columns demonstrates that the distribution of male t-statistics is difficult to distinguish from a draw of randomly generated t-statistics. The maximum value in the third column exceeds the maximum value in the first column, but the first column has more t-statistics clustered above 1.5. In either column, a case can be made for positive treatment effects by focusing on the subset of outcomes near the top. This fact highlights the importance of correcting for multiple testing.

A formal analysis examines summary index FWE p-values and aggregates all long-term outcomes into a single summary index. In comparison to females, each of the nine male summary index coefficients is lower, often by a large margin. Female general effects attain significance for pre-teen, teenage, and adult outcomes in both the Abecedarian and Perry Preschool programs.[44] With the exception of Abecedarian teens, all of these effects at least approach marginal significance after FWE adjustment. Male general effects attain significance only for Abecedarian pre-teens and Early Training Project adults (the latter effect is negative). However, after adjusting for multiple testing, only the Early Training Project adult general effect achieves marginal significance.

A summary test that pools all teen outcomes together across experiments finds an overall effect size of 0.53 for females (standard error of 0.14) and 0.08 for males (standard error of 0.19). The gender difference is statistically significant at the 5 percent level. A summary test that pools all adult outcomes together across experiments finds an overall effect size of 0.27 for females (standard error of 0.09) and -0.05 for males (standard error of 0.11). The gender difference is again statistically

[43]For example, if the Abecedarian high school graduation t-statistic is large, then it is likely that the Abecedarian college attendance t-statistic will also be large. Therefore, patterns of large or small t-statistics are more likely to occur in this data than would be expected in a set of 29 independently sampled t-statistics.

[44]Pre-teen and teenage female general effects are of notable magnitude in the Early Training Project, but do not attain significance because of the limited sample size.

significant at the 5 percent level. Of course, we can never reject an arbitrarily small effect for males, and precision is limited by the relatively small samples. Perhaps real male effects exist but are masked by the standard errors. Nevertheless, the results indicate that any positive male treatment effect is modest at best.

The reported heterogeneity in treatment effects by gender is consistent with several previous findings in the non-experimental literature. For example, Oden, et al. (2000) report that Head Start participation significantly raises high school graduation rates and lowers arrest rates for females. However, no significant effect is found for males. The results also parallel findings in other areas of the human capital literature. Kling and Liebman (2004) report that the Moving to Opportunity program improves educational outcomes and mental health for females, but appears to have *negative* effects on male participants. Abadie, Angrist, and Imbens (2002) find that the Job Training Partnership Act (JTPA) significantly increases female earnings at all quantiles, including a 35 percent increase at the lowest quantile. However, the JTPA has no significant effect on males at any quantile below the median, and the proportional effect never exceeds 12 percent.

A variety of explanations can account for the observed gender differentials. Testing these explanations is beyond the scope of this paper and its data. Nevertheless, a quick summary of possibilities is in order.

One likely possibility is that child development differs between boys and girls. Many researchers believe that girls develop faster than boys. For example, a recent longitudinal study of Australian children found that preschool age females outperform their male counterparts in the physical, social/emotional, and learning domains (Australian Institute of Family Studies, 2005). Evidence is also mounting that education has a greater impact at later stages of development. Fredriksson and Öckert (2005) discover that Swedish children who start school later get more education than their younger peers. This effect is more pronounced for children from weaker socio-economic backgrounds. If additional maturity enhances the effect of schooling, and girls mature faster than boys, then girls should benefit more than boys from early intervention.[45]

Disadvantaged females may also experience different obstacles than disadvantaged males, and

[45]Note that this hypothesis need not be inconsistent with the hypothesis that early intervention is more effective than later intervention. Since free public schooling past age 5 is universally available, later interventions are implicitly being performed on the intensive margin. Early interventions, in contrast, are often performed on the extensive margin. It is therefore possible that early interventions might be more effective, even if education is more effective for older children.

non-cognitive skills developed in preschool might address the obstacles that females face more effectively. A possible example is the role of teen pregnancy in high school dropouts. Males cannot get pregnant, so any effect of preschool on teen pregnancy only benefits females. If teen pregnancy increases the likelihood of dropping out, preschool will have a greater effect on female educational attainment than male educational attainment. However, the data invalidate this particular explanation. Even if pregnancy caused a one-for-one increase in high school dropout status, the observed pregnancy effect still could not explain a majority of the female high school graduation effect. Nevertheless, other differences in obstacles faced by males and females may play important roles.

A third possibility is the existence of a selection effect. "Female" families participating in the program may differ from male families along unobserved dimensions. Gender is typically thought of as randomly assigned, but families with girls may be more or less likely to enroll in preschool programs (the Perry sample, for example, includes significantly more males than females). However, this fact need not invalidate the external validity of the results. If the same selection factors operate in the general population, then the reported female-male differences will be applicable to many preschool programs with voluntary participation.

Finally, recent research has established that students may perform better when taught by teachers of the same gender. For example, Dee (2005) presents evidence that middle school children are perceived as less disruptive and more attentive when the teacher is of the same gender. To my knowledge, all of the preschool teachers in each experiment were female. If preschool age children also perform better when taught by adults of the same sex, then we might expect females to benefit more from early intervention than males.

1.6 Assessing Threats to Validity: Attrition, Violation of Random Assignment, and Clustering

This paper reports significant long-term effects for females in the domains of educational achievement, criminal behavior, marriage, and economic success. The experimental design alleviates concerns about confounding variables, but there remain several issues specific to the individual studies that could cause the treatment and control groups to be systematically different in ways unrelated

to the treatment, or cause statistical tests to over-reject. A careful examination of these issues is necessary before long-term effects for females, and the larger body of experimental preschool research in general, can be readily accepted.

The first problem is attrition, which occurs in the Abecedarian Project and the Perry Preschool Program.[46] The second problem is the intentional exchange of children between groups. This issue occurs only in the Perry Preschool Program. The final issue is clustering, or correlation between individual observations, which occurs primarily in the Perry Preschool Program. I find that the key results remain unchanged after accounting for these problems.

1.6.1 Attrition

Random attrition reduces statistical power but does not cause bias. Non-random attrition is acceptable if it is unrelated to treatment status - it will not induce systematic differences between the treatment and control groups, and estimated effects remain internally valid.[47] Therefore, our only concern is attrition that is caused by assignment status.

Attrition of two types occurs in the preschool experiments. The first type, which I refer to as follow-up attrition, occurs when individuals initially in the sample cannot be located for follow-up interviews, testing, or records collection. The second type, which I refer to as pre-treatment attrition, arises when individuals drop out after receiving their assignment but before entering the sample. In practice, the first type often receives more attention than the second, perhaps because the missing data is readily apparent. Nevertheless, the two types are fundamentally similar.

If attrition is present, the direction of bias it produces is ambiguous. Most of the pre-treatment attrition occurs among children assigned to the treatment groups. In this case, we might expect a positive bias if families that care least about education are the ones refusing treatment.[48] Follow-up attrition affects both treated and control children. The leading causes of follow-up attrition are death and inability to locate the subject. Signing this bias with certainty is infeasible. However,

[46]Only one female is missing for most Early Training Project results. There is no documented evidence of attrition occurring after the random assignment but before data collection. Because attrition is almost non-existent in this study, and because the study found few significant effects to begin with, no attrition analysis is performed for the Early Training Project.

[47]Non-random attrition of this type can still affect the external validity of estimated treatment effects, but this caveat applies to all studies whose participants are not randomly drawn from the relevant population.

[48]On the other hand, it is possible that some treatment families pulled their children out because they felt they could offer a better experience at home. Depending on the characteristics of these families, this explanation could lead to a negative bias.

it is notable that more control children died than treated children. If control children who die are especially poor or disadvantaged, attrition from death would attenuate a positive treatment effect. If successful subjects are likely to move out of state, then attrition from movement would also attenuate a positive treatment effect. We therefore might expect follow-up attrition to exert a negative bias. We cannot accurately guess the direction of the overall bias.

Abecedarian

The Abecedarian Project lost eleven children to pre-treatment attrition. Seven treatment group families and one control group family withdrew upon receiving their group assignments. Two control group children received preschool treatment due to medical conditions requiring close supervision; these children are not present in the dataset. An additional seven children were lost to follow-up attrition for most outcomes. One treatment male, one treatment female, and two control females died early in life. Three additional subjects are not present for various reasons: one treatment female had a seizure disorder, one control female withdrew for family related matters, and one treatment female declined to participate in the age 21 interview.[49]

Table 1.11 reports estimates for key outcomes under a variety of attrition assumptions. The analysis focuses on females, because males suffer less attrition and demonstrate no significant effects.[50] Columns (1) and (2) focus on follow-up attrition only. Of the six missing females, four dropped out for medical reasons unlikely to be affected by treatment status (three deaths and one seizure disorder).[51] The analysis therefore explores imputations for the two females that specifically chose not to participate in follow-up surveys.[52] Column (1) assigns the missing treated female the

[49]The information regarding attrition comes from Campbell, et al. (2002), Clarke and Campbell (1998), Campbell and Ramey (1994), and Ramey, Yeates, and Short (1984).

[50]Under extreme assumptions regarding missing values, some results for males could attain marginal significance. Such results would not constitute compelling evidence of a male treatment effect.

[51]All of these deaths occurred at an early age - generally less than one year. One might hypothesize that preschool affects infant death rates, particularly those resulting from accidents or infectious diseases. CDC data indicates that the magnitude of this effect would be trivial. Of the top causes of black postneonatal death in 1979 - which account for almost 70 percent of total black postneonatal deaths - preschool attendance could only affect accidents, homicide, pneumonia, bronchitis, viral infections, and meningitis to a significant degree. These causes account for only 19 percent of postneonatal deaths (Hoyert, Kochanek, and Murphy, 1999). Even if preschool induced a 50 percent change in death rates from these causes, total death rates would change by only 9.5 percent. The death rates in the actual sample match this prediction: two treatment group and two control group children died. Of course, it is theoretically possible that preschool could prevent some deaths and cause others, so dramatic effects for particular causes could be masked at the aggregate level. This seems unlikely.

[52]There is no high school graduation information for one additional treated female. Therefore, relative to the other results, the high school graduation results assign values for one additional treated female.

25th percentile of each variable and the missing control female the 75th percentile of each variable (for all variables, higher percentiles correspond to "better" outcomes). Column (2) assigns the missing treated female the 10th percentile of each variable and the missing control female the 90th percentile of each variable. In both columns, the two significant outcomes - college attendance and marijuana use - remain significant.

Columns (3) through (6) address both follow-up and pre-treatment attrition. Column (3) assigns missing values as follows: the missing follow-up treated subject receives the 25th percentile for each variable, the missing follow-up control subject receives the 75th percentile for each variable, four of the missing pre-treatment treated subjects receive the 25th percentile for each variable, and two of the missing pre-treatment control subjects receive the 75th percentile for each variable.[53] Column (4) is identical to column (3) except that the missing follow-up subjects are assigned the 10th and 90th percentiles respectively. Column (5) is identical to column (4) except that the missing pre-treatment subjects are assigned the 10th and 90th percentiles respectively. Column (6) implements the "worst case" scenario: all attrition is assumed non-random, all missing subjects are assumed female unless otherwise identified, all missing treated subjects are assigned the 10th percentile values, and all missing control subjects are assigned the 90th percentile values. The worst case scenario assigns values to a total of seventeen missing subjects.

The results in columns (3) through (6) demonstrate that some Abecedarian effects retain significance under all but extreme assumptions about missing values. Both college attendance and marijuana use remain significant in columns (3) and (4). These variables lose significance in column (5), when six missing pre-treatment subjects are assigned values at the 10th and 90th percentiles of the distribution. The coefficients approach zero in column (6) under the worst case scenario; however, the assumptions underlying this scenario are implausible.

Perry

The Perry Preschool Project lost five children to pre-treatment attrition. Four treatment group children moved away before completing preschool, and one control group child died (Schweinhart, et al., 2005). None of these children entered the dataset. However, for several key measures, there

[53] A total of seven treated subjects and four control subjects are missing from pre-treatment attrition. I do not have information on their genders, so in the base case I assign genders to the missing pre-treatment subjects based on the gender distribution of the non-missing sample.

is no follow-up attrition.

Table 1.12 presents estimates for key outcomes under three sets of assumptions. As with Abecedarian, the analysis focuses on females. The pre-treatment attrition in Perry is plausibly independent of treatment status. 80 percent of the pre-treatment attrition occurred when four treatment children moved away before completing the program. No control child moved away during the same period, and it is doubtful that the offer of free schooling would make a family *more* likely to leave the area. This attrition is therefore unlikely to be related to treatment status. An additional control child died at an early age and was not included in the sample. This death is unlikely to be the result of treatment status.[54]

Columns (1) and (2) in Table 1.12 address follow-up attrition only. Since marital status, high school graduation status, and government transfer data are available for all individuals, the reported coefficients for these variables are identical to the original results. Column (1) assigns missing treated subjects the 25th percentile of each variable conditional on high school graduation status. It assigns missing control subjects the 75th percentile of each variable conditional on high school graduation status.[55] All variables remain significant in column (1). Column (2) is identical to column (1) except that the 25th and 75th percentiles are replaced with the 10th and 90th percentiles respectively. Every variable except employment remains significant. Column (3) implements the "worst case" scenario. For variables with follow-up attrition, column (3) assigns missing treated subjects the 10th percentile of each variable conditional on high school graduation status and missing control subjects the 90th percentile of each variable conditional on high school graduation status. The four treated subjects that moved away are assumed to be female and assigned the 10th percentile of each variable. The one dead control subject is assumed to be female and assigned the 90th percentile of each variable. This worst case scenario eliminates the significance of most variables. However, the high school graduation effect remains significant despite the extreme assumptions underlying this scenario.

[54]Please see the note in Section 1.6.1 regarding attrition due to death.

[55]This procedure leverages information contained in the complete high school graduation data for predictive purposes.

1.6.2 Violation of Random Assignment

For the most part, families complied with their initial group assignments. Those that refused were generally dropped from the data, as described in Section 1.6.1. However, in the Perry Preschool Project, several children with working mothers were exchanged with select control group children. Two of these switches may have occurred without replacement. The exchanges were made because the employed mothers could not accommodate the program's weekly home visits. Replacement children were purportedly matched on initial IQ, but confounding may still occur because maternal presence at an early age could affect later outcomes.

In no case did the Perry researchers record original assignment status. This fact precludes the use of an instrumental variables approach, so I perform alternative tests to gauge the impact of these violations. First, I condition on initial maternal employment status. However, five children with employed mothers were not transfered from the treatment group. Conditioning upon maternal employment status is therefore insufficient, because children with employed mothers who switched may differ in important ways from those with employed mothers who stayed. Furthermore, conditioning on maternal employment does not account for the control children who were exchanged to the treatment group. If these children were matched with the maternal employment children, they may differ from the average child in expectation. To address these issues, I examine a range of possible group assignments and the corresponding coefficient estimates.

The first two columns of Table 1.13 present results for key outcomes for both genders controlling for the effect of maternal employment at entry. These results do not differ markedly from the original estimates. In fact, the coefficients are of slightly greater magnitude after controlling for maternal employment. All female effects remain significant, and one male effect - lifetime arrests at 27 - achieves marginal significance.

I conduct further analysis for key female outcomes under the assumption that four treatment children with employed mothers were switched with four control children without employed mothers. Records indicate that either two or five children with employed mothers switched from the treatment group, but probability estimates indicate that the number could have been as high as eight (Schweinhart, et al., 2005). There is no record of the gender distribution of exchanges. However, the data suggest that approximately three females switched from the treatment group, since there are nine control females with employed mothers as compared to three treated females with

employed mothers. The assumption that four treated females were exchanged is therefore likely to overestimate the total number of female exchanges.

The exchange analysis examines every possible combination of switches that swaps four treated females with employed mothers for four control females without employed mothers. The number of possible combinations totals 921,600. For each combination, I estimate the treatment effect for six key outcomes using instrumental variables. The hypothesized original group assignment serves as the instrument.

Because each individual carries an entire set of outcomes, it is meaningless to tabulate the resulting t-statistics in isolation. In order to compare different combinations of exchanges, I construct an average t-statistic for each combination equal to the mean of the six estimated t-statistics.[56] I then rank combinations according to their average t-statistics.

The last five columns of Table 1.13 report female results for different quantiles of the average t-statistic. The first column presents results at the median of the distribution, the second at the 25th quantile, the third at the 10th quantile, the fourth at the 1st quantile, and the fifth at the distribution's minimum value. At the median, the coefficients are of similar magnitude to the original OLS results, but the standard errors have increased because the instrument is not perfectly correlated with treatment status. Consequently, some results are insignificant, but the two arrest variables and the graduation variable remain significant. At the 10th quantile, the graduation and arrest outcomes attain marginal significance. At the bottom of the distribution they are all insignificant. However, these quantiles are identified ex post. When the Perry researchers chose which control individuals to exchange with treated individuals, they could only guess at future outcomes. Even if the researchers tried make exchanges that would benefit the treatment group and hurt the control group (an implausible assumption), it is unlikely they could achieve that goal to as great a degree as implied in the 10th or 1st quantile columns. The last three columns of Table 1.13 therefore correspond to very extreme assumptions, but even in the 1st quantile column one coefficient remains statistically significant. It is therefore unlikely that exchanges based on maternal employment status drive the significance of the results.

[56] When constructing the average t-statistic, I reverse the sign on the two arrest t-statistics and the government transfer t-statistic, so that positive t-statistics always correspond to "better" outcomes.

1.6.3 Clustering

The p-values presented in Section (1.4) are robust to distributional assumptions.[57] Nevertheless, clustering issues could bias the standard errors, causing conventional tests to overstate the significance of the results, particularly in the case of the Perry Preschool Project.

It is well established that clustering - or correlation across observations - can severely inflate test statistics if not properly accounted for (Bertrand, Duflo, and Mullainathan, 2004). In these experiments, there are two possible sources of interdependence that could be correlated within treatment status groups.[58] First, peer or class effects might lead to correlations between students within a given preschool class.[59] Second, the automatic assignment of younger siblings to the same treatment group as their older siblings reduces the number of independent observations.

Previous research has demonstrated that negative peer effects can lower class achievement (for example, Figlio, 2005). It is therefore plausible that a poorly behaved child may reduce the performance of her preschool peers, implying an intra-class correlation.[60] Furthermore, within each class the treatment variable is perfectly correlated. The standard errors could therefore be too small because children within each class are mistakenly treated as independent observations.

To address this problem, I collapse the data down to cohort-by-treatment status means. For the Perry females, with five cohorts and two treatment statuses, this procedure reduces the dataset to ten observations. I estimate an OLS regression using these ten observations. The first row of Table 1.14 presents the results. Despite the small sample, five of the six key variables remain statistically significant.[61] The only outcome that loses significance is the employment variable. I cannot run a similar regression for the Abecedarian children as I do not have their cohort identification data, but the Perry analysis suggest that intra-class clustering does not drive the significance of the results.

[57] These p-values do not differ markedly from the p-values generated by conventional t-tests, so standard OLS t-tests are presented for the remaining results.

[58] Inflation only occurs when there is a similar correlation structure in both the dependent and independent variables. For example, the fact that cohorts might face similar shocks will not bias the standard errors because treatment status is randomly assigned within a given cohort.

[59] Angrist and Lang (2004), for example, find evidence of negative peer effects in the Boston METCO program.

[60] Peer effects may operate more strongly for poorly behaved children than for well behaved children. In that case, they will tend to reduce the estimated effect. However, this is not a source of coefficient bias, since it is a direct consequence of the preschool program. Rather, it only effects the external validity of the results when applied to different demographic groups.

[61] Donald and Lang (2001) argue that t-statistics can be misleading when the number of groups - in this case, cohort-by-treatment status units - is small, because common shocks may not be normally distributed. However, for these results I conducted simulations drawing common group effects from a distribution with all its probability weight at the tails. These simulations did not generate poorly distributed t-statistics.

Another clustering problem arises from the assignment of younger siblings to the same treatment status as their older siblings. Performance is almost surely correlated within families, and this assignment mechanism guarantees that treatment status is also correlated within families. To address this problem, I restrict the sample to eldest siblings and only children. For Perry females, this restriction decreases the sample size from 51 to 37. The results for Perry females are reported in the second row of Table 1.14. All presented outcomes remain strongly significant. In the Abecedarian program, the sample contains only two sibling pairs, so an older sibling analysis is unnecessary.

A final clustering issue is the possibility of teacher effects.[62] Individual level data on teacher assignment is not available. However, Perry Preschool employed ten teachers, the Early Training Project employed two teachers and several assistants, and the Abecedarian Project employed multiple teachers (the exact number is unclear). It is therefore unlikely that the observed effects are the result of one or two stellar teachers.

1.7 Conclusion

This paper conducts a robust reanalysis of the influential experimental preschool literature. It partially confirms previous findings, presenting strong evidence that females benefit from early intervention. Significant female effects appear in the domains of criminal behavior, marriage, and economic success, but the most consistent improvement is an increase in total years of schooling. These results are robust to reasonable concerns regarding attrition, violation of random assignment, and clustering. Many female results also remain significant after adjusting for multiple inference.

For males, however, there is no evidence of positive, long-term preschool treatment effects. Most coefficients are insignificant, and several of the significant coefficients imply an adverse effect. The overall pattern of male coefficients is consistent with the hypothesis of a minimal treatment effect at best - significant effects go in both directions and appear at a frequency one would expect simply due to chance.[63]

The observed differences between female and male treatment effects are significant in a number of cases, particularly with respect to total years of education. The difference in overall long-term

[62]Technically, teacher effects would be an issue of external validity, not internal validity.

[63]Previous research has overlooked this result because there has been no systematic analysis by gender across experiments, nor has anyone applied a unified statistical framework that is robust to problems of multiple inference.

female and male summary indices is also significant. However, additional research with new data is necessary to determine the exact magnitude of the female-male treatment effect differential, and to discover whether males derive modest benefits from preschool intervention or no benefits at all.

In the context of the current human capital literature, this paper makes clear several points. Foremost, intensive preschool intervention does positively affect later life outcomes, at least for disadvantaged African-American females. However, there is no evidence of strong long-term preschool benefits for males. This fact suggests that investments in early education alone may not dramatically improve opportunities for disadvantaged males. The indicated treatment effect heterogeneity also calls into question the external applicability of experimental estimates. If treatment effects vary by gender, it is plausible that they may also vary by race or class. Richer variation in sample demographics is necessary for the design of optimal human capital policy. As Hanushek (2003) suggests, financing broader experimental research on human capital investments may well yield the highest return today of any human capital policy.

Figure 1-1: Effects of Preschool on Teen and Adult Outcomes

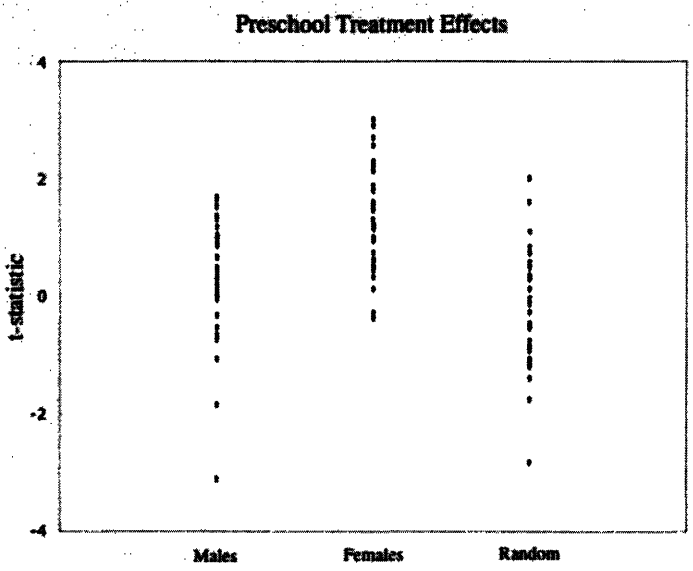

Table 1.1: Summary Statistics

Variable	Abecedarian	Perry	Early Training
Percent treated	51.4	47.2	67.7
	(50.2)	(50.1)	(47.1)
Percent female	53.2	41.5	46.2
	(50.1)	(49.5)	(50.2)
IQ age 5	97.8	88.9	91.5
	(12.6)	(12.9)	(13.6)
IQ age 14-17	93.2	80.9	77.7
	(10.3)	(11.0)	(13.2)
Percent retained in grade	45.6	37.5	54.2
	(50.1)	(48.6)	(50.2)
Percent graduate HS	69.9	61.8	60.0
	(46.1)	(48.8)	(49.4)
Percent employed as adult	57.3	62.1	N/A
	(49.7)	(48.7)	
Percent with criminal record	43.3	52.8	N/A
	(49.8)	(50.1)	

Notes: Parentheses contain standard deviations.

Table 1.2: Effects on Pre-Teen IQ Scores

			Female				Male				Gender Interaction
Outcome	Age	Project	**Effect**	CM	p-val	N	**Effect**	CM	p-val	N	t-stat
IQ	5	ABC	4.94	96.76	0.182	48	10.19	90.81	0.005	47	-1.05
			(3.58)				(3.52)				
IQ	6.5	ABC	5.13	92.96	0.135	46	7.18	92.10	0.058	45	-0.41
			(3.35)				(3.65)				
IQ	12	ABC	8.35	87.35	0.004	52	3.21	90.48	0.291	49	1.24
			(2.75)				(3.10)				
IQ	5	Perry	12.67	81.65	0.004	39	10.61	84.79	0.000	54	0.40
			(4.30)				(2.84)				
IQ	6	Perry	3.75	87.16	0.243	48	5.66	85.82	0.037	72	-0.46
			(3.21)				(2.68)				
IQ	10	Perry	4.96	81.79	0.169	43	-2.33	86.03	0.375	71	1.70
			(3.45)				(2.56)				
IQ	5	ETP	13.55	87.60	0.018	30	4.43	87.18	0.232	34	1.28
			(6.09)				(3.75)				
IQ	7	ETP	8.61	89.89	0.119	29	4.11	92.89	0.346	30	0.57
			(6.69)				(4.25)				
IQ	10	ETP	9.79	81.56	0.069	29	-3.17	88.33	0.505	27	1.68
			(5.73)				(5.15)				

Notes: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. P-values are computed as described in Section (1.3); t-statistics test the difference between female and male treatment effects.

Table 1.3: Effects on Pre-Teen Primary School Outcomes

			Female				Male				Gender Interaction
Outcome	Age	Project	**Effect**	CM	p-val	N	**Effect**	CM	p-val	N	t-stat
Retained	12	ABC	-0.229	0.429	0.082	53	-0.188	0.545	0.201	50	-0.21
			(0.125)				(0.142)				
Spec Educ	12	ABC	-0.066	0.296	0.576	53	-0.269	0.591	0.054	50	1.10
			(0.123)				(0.140)				
Repeat Grade	12	Perry	-0.201	0.409	0.134	46	0.078	0.389	0.514	66	-1.51
			(0.137)				(0.124)				
Spec Educ	17	Perry	-0.262	0.462	0.060	51	-0.037	0.462	0.741	72	-1.28
			(0.129)				(0.119)				
Retained	17	ETP	-0.284	0.600	0.156	29	0.100	0.600	0.514	30	-1.40
			(0.195)				(0.192)				
Special Help	17	ETP	0.116	0.200	0.529	29	0.036	0.364	0.832	31	0.31
			(0.171)				(0.188)				

Notes: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. P-values are computed as described in Section (1.3); t-statistics test the difference between female and male treatment effects.

Table 1.4: Summary Index Effects

		Female			Male			Gender Interaction
Project	Age	**Effect**	FWE p-val	N	**Effect**	FWE p-val	N	t-stat
ABC	Pre-Teen	0.445	0.117	54	0.417	0.187	51	0.11
		(0.194)			(0.181)			
Perry	Pre-Teen	0.537	0.026	51	0.150	0.941	72	1.53
		(0.177)			(0.172)			
ETP	Pre-Teen	0.380	0.346	30	0.142	0.960	34	0.67
		(0.270)			(0.238)			
ABC	Teen	0.422	0.153	53	0.162	0.941	51	0.93
		(0.202)			(0.194)			
Perry	Teen	0.613	0.003	51	0.035	0.976	72	3.32
		(0.156)			(0.096)			
ETP	Teen	0.551	0.346	29	0.097	0.976	32	0.95
		(0.327)			(0.345)			
ABC	Adult	0.452	0.022	53	0.312	0.369	51	0.64
		(0.144)			(0.166)			
Perry	Adult	0.358	0.117	51	-0.017	0.976	72	1.88
		(0.151)			(0.130)			
ETP	Adult	-0.067	0.709	29	-0.654	0.090	31	1.82
		(0.188)			(0.257)			

Notes: Parentheses contain OLS standard errors. FWE p-values are computed as described in Section (1.3); t-statistics test the difference between female and male treatment effects. See Table 1.5 for the components of each summary index.

Table 1.5: Summary Index Components

Project	Stage	Summary Index Components
ABC	Pre-Teen	IQ (5, 6.5, 12), Retained in Grade (12), Special Education (12)
Perry	Pre-Teen	IQ (5, 6, 10), Repeat Grade (17), Special Education (17)
ETP	Pre-Teen	IQ (5, 7, 10), Retained in Grade (17), Special Help (17)
ABC	Teen	IQ (15), HS Grad (18), Teen Parent (19)
Perry	Teen	IQ (14), HS Grad (18), Unemployed (19), Transfers (19), Teen Parent (19), Arrested (19)
ETP	Teen	IQ (17), HS Drop Out (18), Worked (18)
ABC	Adult	College (21), Employed (21), Convicted (21), Felon (21), Jailed (21), Marijuana (21)
Perry	Adult	College (27), Employed (27, 40), Income (27, 40), Criminal Record (27), Arrests (27), Drugs (27), Married (27)
ETP	Adult	College (21), Receive Income (21), On Welfare (21)

Notes: Age of measurement in parentheses.

Table 1.6: Effects on Teenage Academic Outcomes

			Female				Male				Gender Interaction
Outcome	Age	Project	**Effect**	CM	p-val	N	**Effect**	CM	p-val	N	t-stat
IQ	15	ABC	4.22	89.50	0.142	53	4.66	92.48	0.091	51	-0.11
			(2.85)				(2.79)				
IQ	14	Perry	2.64	76.77	0.313	46	-0.96	83.26	0.761	64	0.91
			(2.57)				(3.03)				
IQ	17	ETP	2.08	76.11	0.744	25	1.64	76.78	0.737	28	0.05
			(6.80)				(5.09)				
HS Grad	18	ABC	0.226	0.607	0.086	52	-0.096	0.739	0.465	51	1.80
			(0.122)				(0.131)				
HS Grad	18	Perry	0.494	0.346	0.000	51	-0.061	0.667	0.583	72	3.32
			(0.121)				(0.115)				
Ever Drop Out of HS	18	ETP	-0.289	0.500	0.107	29	-0.095	0.545	0.676	31	-0.72
			(0.190)				(0.193)				

Notes: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. P-values are computed as described in Section (1.3); t-statistics test the difference between female and male treatment effects.

Table 1.7: Effects on Teenage Economic and Social Outcomes

			Female				Male				Gender Interaction
Outcome	Age	Project	**Effect**	CM	p-val	N	**Effect**	CM	p-val	N	t-stat
Unemp	19	Perry	-0.308	0.708	0.028	49	-0.021	0.385	0.877	72	-1.60
			(0.138)				(0.116)				
Transfers	19	Perry	-1,569	2,828	0.035	51	-28	398	0.933	72	-1.96
			(722)				(319)				
Ever Work	18	ETP	0.125	0.500	0.581	22	-0.063	1.000	0.641	23	0.73
			(0.249)				(0.063)				
Teen Parent	19	ABC	-0.211	0.571	0.133	53	-0.126	0.304	0.315	51	-0.47
			(0.137)				(0.123)				
Had Child	19	Perry	-0.187	0.667	0.209	49	-0.044	0.256	0.666	72	-0.82
			(0.142)				(0.101)				
Arrested	19	Perry	-0.337	0.417	0.006	49	-0.079	0.564	0.527	72	-1.54
			(0.117)				(0.119)				

Notes: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. P-values are computed as described in Section (1.3); t-statistics test the difference between female and male treatment effects.

Table 1.8: Effects on Adult Academic Outcomes

			Female				Male				Gender Interaction
Outcome	Age	Project	**Effect**	CM	p-val	N	**Effect**	CM	p-val	N	t-stat
In College	21	ABC	0.293	0.107	0.015	53	0.148	0.174	0.258	51	0.87
			(0.116)				(0.121)				
Any College	27	Perry	0.160	0.280	0.256	50	-0.005	0.308	0.978	72	0.94
			(0.137)				(0.110)				
In Post HS Educ	21	ETP	0.121	0.300	0.537	29	-0.486	0.636	0.005	31	2.37
			(0.191)				(0.171)				

Notes: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. P-values are computed as described in Section (1.3); t-statistics test the difference between female and male treatment effects.

Table 1.9: Effects on Adult Economic Outcomes

			Female				Male				Gender Interaction
Outcome	Age	Project	**Effect**	CM	p-val	N	**Effect**	CM	p-val	N	t-stat
Employed	21	ABC	0.104	0.536	0.432	53	0.188	0.455	0.196	50	-0.43
			(0.137)				(0.142)				
Employed	27	Perry	0.255	0.545	0.076	47	0.036	0.564	0.765	69	1.20
			(0.136)				(0.121)				
Annual Income	27	Perry	2,567	8,986	0.353	47	2,363	12,495	0.382	66	0.05
			(2,686)				(2,708)				
Employed	40	Perry	0.015	0.818	0.922	46	0.200	0.500	0.109	66	-1.12
			(0.115)				(0.120)				
Annual Income	40	Perry	3,492	17,374	0.536	46	6,228	21,119	0.302	66	-0.34
			(5,491)				(5,958)				
Receive Income	21	ETP	-0.074	0.600	0.688	29	-0.159	0.909	0.303	31	0.36
			(0.200)				(0.134)				
Receive Welfare	21	ETP	-0.042	0.200	0.805	30	N/A	0.000	N/A	35	N/A
			(0.157)				(N/A)				

Notes: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. P-values are computed as described in Section (1.3); t-statistics test the difference between female and male treatment effects.

Table 1.10: Effects on Adult Social Outcomes

			Female				Male				Gender Interaction
Outcome	Age	Project	**Effect**	CM	p-val	N	**Effect**	CM	p-val	N	t-stat
Convicted	21	ABC	-0.101	0.143	0.224	52	-0.089	0.348	0.523	50	-0.08
			(0.079)				(0.133)				
Felony	21	ABC	N/A	0.000	N/A	52	-0.113	0.261	0.369	50	N/A
			N/A				(0.117)				
Jailed	21	ABC	-0.030	0.071	0.703	52	-0.177	0.391	0.160	51	1.01
			(0.065)				(0.131)				
Marijuana User	21	ABC	-0.317	0.357	0.003	53	-0.127	0.435	0.390	49	-1.10
			(0.101)				(0.140)				
Criminal Record	27	Perry	-0.146	0.346	0.260	51	-0.021	0.718	0.824	72	-0.75
			(0.125)				(0.109)				
Lifetime Arrests	27	Perry	-1.95	2.27	0.012	49	-2.31	6.10	0.133	72	0.21
			(0.83)				(1.50)				
Ever Used Drugs	27	Perry	-0.157	0.300	0.213	41	0.198	0.189	0.073	68	-2.08
			(0.131)				(0.110)				
Married	27	Perry	0.317	0.083	0.008	49	0.002	0.256	0.983	70	2.01
			(0.115)				(0.107)				

Notes: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. P-values are computed as described in Section (1.3); t-statistics test the difference between female and male treatment effects.

Table 1.11: Attrition Analysis for Key Abecedarian Variables

Outcome	Age	(1)	(2)	(3)	(4)	(5)	(6)
High School Grad	18	0.149	0.149	0.022	0.022	0.022	-0.061
		(0.125)	(0.125)	(0.124)	(0.124)	(0.124)	(0.120)
Attending College	21	0.281	0.247	0.237	0.204	0.140	0.061
		(0.110)	(0.115)	(0.102)	(0.106)	(0.113)	(0.111)
Marijuana User	21	-0.306	-0.268	-0.289	-0.256	-0.123	-0.030
		(0.102)	(0.108)	(0.093)	(0.098)	(0.113)	(0.113)
Teen Parent	19	-0.167	-0.167	-0.049	-0.049	-0.049	0.030
		(0.135)	(0.135)	(0.130)	(0.130)	(0.130)	(0.125)

Notes: Parentheses contain OLS standard errors.

Table 1.12: Attrition Analysis for Key Perry Variables

Outcome	**Age**	**(1)**	**(2)**	**(3)**
Employed	27	0.300	0.185	0.071
		(0.130)	(0.128)	(0.127)
Married	27	0.285	0.285	0.179
		(0.118)	(0.118)	(0.118)
Ever Arrested	19	-0.305	-0.305	-0.179
		(0.113)	(0.113)	(0.118)
High School Grad	18	0.494	0.494	0.357
		(0.122)	(0.122)	(0.126)
Transfers	19	-1569	-1569	-945
		(729)	(729)	(716)
Lifetime Arrests	27	-1.95	-1.95	-1.39
		(0.84)	(0.84)	(0.81)

Notes: Parentheses contain OLS standard errors.

Table 1.13: Effects of Maternal Employment on Key Perry Results

		Control for WM		Alternative Assumptions on WM Swaps				
Outcome	Age	Female	Male	50th	25th	10th	1st	Lowest
Employed	27	0.316	0.115	0.225	0.244	0.048	-0.003	-0.003
		(0.141)	(0.122)	(0.274)	(0.234)	(0.260)	(0.243)	(0.243)
Married	27	0.318	0.039	0.306	0.192	0.391	0.259	-0.017
		(0.123)	(0.111)	(0.226)	(0.199)	(0.213)	(0.198)	(0.214)
Ever Arrested	19	-0.398	-0.083	-0.661	-0.500	-0.352	-0.431	-0.293
		(0.118)	(0.126)	(0.242)	(0.199)	(0.210)	(0.197)	(0.195)
High School Grad	18	0.530	-0.005	0.581	0.530	0.373	0.296	0.296
		(0.126)	(0.119)	(0.228)	(0.200)	(0.216)	(0.206)	(0.206)
Transfers	19	-1765	-144	-2254	-1078	-2102	-794	-670
		(756)	(337)	(1368)	(1200)	(1287)	(1213)	(1217)
Lifetime Arrests	27	-2.30	-2.90	-3.70	-2.77	-2.78	-2.30	-2.17
		(0.86)	(1.56)	(1.64)	(1.40)	(1.50)	(1.39)	(1.39)

Notes: WM = Working Mothers. Results under alternative assumptions are estimated using hypothesized group assignment as an instrument for treatment status. Parentheses contain OLS standard errors when controlling for working mothers, and IV standard errors when examining results under alternative assumptions about the working mother swaps. Sample size varies within columns due to attrition for some variables.

Table 1.14: Effects of Clustering on Key Perry Results

Model	Employed at 27	Married at 27	Arrested by 19	High School Graduate	Transfers at 19	Lifetime Arrests at 27
Collapsed to Cohort	0.195	0.336	-0.303	0.477	-1703	-1.59
by Treatment Means	(0.178)	(0.171)	(0.136)	(0.166)	(889)	(0.74)
Eldest Siblings and	0.342	0.409	-0.307	0.561	-2563	-1.85
Only Children Sample	(0.151)	(0.146)	(0.139)	(0.134)	(859)	(0.88)

Notes: For results estimated using cohort by treatment means, N=10. Parentheses contain OLS standard errors.

1.8 Stata Pseudo-Code

Sample Stata code for the free step-down resampling algorithm follows. Some code has been changed to improve readability and would not run as literally written. This code is adapted from Algorithm 2.8 in Westfall and Young (1993).

```
local counter = 1
* run the original regressions and create the p-val simulation storage counters
foreach lhsvar in outcome-varlist {
regress 'lhsvar' treated
replace t-stat = abs(_b[treated]/_se[treated]) in 'counter'
replace p-val = 2*ttail(e(N),t-stat)
local 'lhsvar'-count = 0
local counter = 'counter' + 1
}

* sort the regressions according to ascending p-value.
sort p-val
sort outcome-varlist by ascending p-val

* store the total number of tests originally conducted
local endvar = 'counter' - 1
* initialize the simulation counter
local iteration = 1

* run 10,000 iterations of the simulation; record results in p-val storage counters
while 'iteration' ¡= 10000 {
replace simtreatment-uni = uniform()
replace simtreatment = (simtreatment-uni > 0.5) if perry==1 or abc==1
replace simtreatment = (simtreatment-uni > 0.67) if etp==1
local counter = 1
```

```
foreach lhsvar of outcome-varlist {
regress 'lhsvar' simtreatment
replace t-stat-sim = abs(_b[simtreatment]/_se[simtreatment]) in 'counter'
replace p-val-sim = 2*ttail(e(N),t-stat-sim) in 'counter'
local counter = 'counter' + 1
}
```

* enforce monotonicity in the simulated p-vals and then tabulate whether the simulated p-vals are less than the respective original p-vals

```
local countdown = 'endvar'
foreach lhsvar of reverse-outcome-varlist {
replace p-val-sim = min(p-val-sim, p-val-sim[_n+1]) in 'countdown'
if p-val-sim['countdown'] <= p-val['countdown'] {
local 'lhsvar'-count = 'lhsvar'-count + 1
}
local countdown = 'countdown' - 1
}
local iteration = 'iteration' + 1
}
```

* calculate the adjusted p-val as the ratio of the number of times that the simulated p-vals are less than the original p-val divided by the total number of iterations; enforce the ordering of the original p-values

```
local counter = 1
foreach lhsvar of outcome-varlist {
replace p-vals = max(round('lhsvar'-count/10000, .001), p-vals['counter'-1]) in 'counter'
local counter = 'counter' + 1
}
```

Chapter 2

The Effects of Social Status on Health: Evidence from Whitehall

Coauthored with Professor Sir Michael Marmot

2.1 Introduction

A long-standing debate exists among economists regarding the relationship between socioeconomic status and health. The positive cross-sectional correlation between social status and health is well established (Marmot, 2003). Nevertheless, some broad surveys of the literature document the difficulty in measuring the causal effect of social status on health as defined by Rubin (1974) and summarize the empirical evidence as inconclusive (Smith, 1999; Deaton, 2003). Some research suggests that lagged social status predicts future health outcomes (for example, Adda, Chandola, and Marmot, 2003), and a large body of research examines the structural channels through which the observed health gradient may operate (for example, Marmot, et al., 1997, Kuper and Marmot, 2003, and Chandola, Siegrist, and Marmot, 2005). However, there is little evidence on the effect of an experimental manipulation of social status on health outcomes (Mealli and Rubin, 2003). It is also difficult to disentangle the separate effects of education and income, both of which are important determinants of social status (Deaton and Paxson, 1999).

One salient finding in the literature is that income differentials across developed countries do not affect life expectancy, but income differentials within developed countries are strongly related

to health (Deaton, 2003). Deaton and Paxson (1999) present a framework for understanding these patterns. Within this framework, health is an increasing function of the difference between an individual's income and the average income of his or her reference group. As a result, the independent variable of interest - the gap between observed income and average reference group income - is unobserved. Instead, a noisy measure of this variable is observed. Conventional linear regressions of health on income may therefore understate the health effects of increased income. This model explains the divergence between cross-country and within-country health gradients, and it suggests a way to minimize the attenuation bias.[1] Specifically, if we limit the study sample to a single reference group, then within that reference group observed income will be perfectly correlated with the variable of interest.

The Whitehall II data focus on a plausible "ready-made" reference group: British civil servants working in Inner London. The original Whitehall study collected data on over 18,000 white-collar male civil servants in the vicinity of the Whitehall area of London (Reid, et al., 1974). Although it was not the original aim of the study, Marmot, et al. (1978) documented a significant relationship between employment grade and coronary heart disease. This finding surprised many researchers, who did not expect to observe a large health gradient in a relatively homogenous population. The substantial body of research originating from the first Whitehall study aided the design of a second Whitehall study, devised specifically to explore the causal channels between employment grade and health. This study, known as Whitehall II, collected longitudinal data on over 10,000 white-collar civil servants beginning in 1985. It reconfirmed the original Whitehall results and advanced the hypothesis that factors such as increased job control could underlie the relationship between employment grade and heart disease (Marmot, et al., 1991; Bosma, et al., 1997). However, it has proven difficult to produce a convincing estimate of the effect of an experimental manipulation of employment grade on coronary heart disease. Many economists have therefore been reluctant to interpret the Whitehall results as measuring the causal effects of employment grade on health outcomes (Smith, 1999).

To measure the causal effect of promotions on coronary heart disease, we use a plausibly random

[1]It is important to emphasize that cross-country health gradient is the relationship between average income and health outcomes across countries. A large literature, summarized in Wilkinson and Picket (2005), also exists on the relationship between income inequality and health outcomes across countries. Since this literature uses income dispersion within countries as the independent variable, it is more relevant to the within-country gradient than the cross-country gradient.

source of variation in employment grade. Specifically, we exploit variation in promotion rates across major Civil Service departments. For any given promotion slot, candidates are selected on the basis of merit (Stanley, 2004). However, within departments candidates can only be promoted if a slot is available, and departments cannot easily change the number of available promotion slots. Departmental promotion rates therefore have little relationship to average civil servant quality within a department after conditioning upon grade of entry. Instead, they are a complex function of relative cohort sizes, departmental grade composition, and overall employee departure rates (HM Treasury, 1985). Our empirical strategy therefore uses between-department promotion rates as an instrument to identify the effects of promotion on coronary heart disease.

Of course, it is possible that civil servants could select into departments based on future expectations of promotion rates. If these expectations are correlated with actual promotion rates, our instrument may not be valid. We therefore analyze a range of observable health measures that should not be affected by promotion to establish whether pre-treatment health is correlated with departmental promotion rates. Overall, we find that the correlation between pre-treatment health characteristics and departmental promotion rates is insignificant and an order of magnitude smaller than our instrumental variables estimates for heart disease. We also demonstrate that, for variables for which we have consistent, repeated measures over time - self-reported health and chest pain - the estimated effect appears to be increasing in time since promotion. This fact provides further evidence suggesting that our instrumental variables model is estimating a causal effect rather than a selection effect. To complement the coronary heart disease results, we also examine a variety of other outcomes, including self-reported health, mortality rates, self-reported heart trouble, SF-36 survey physical and mental health component scores, and marital rates.

The results suggest that promotions can reduce the probability of heart disease by 3 to 13 percentage points over a 15 year period, and may improve other aspects of physical health as well. However, we find no evidence that promotions improve mental health. The coronary heart disease estimates are several times larger than cross-sectional estimates in the previous Whitehall literature, but they are consistent with other research on the causal effects of social status on health.

The paper is organized as follows. Section (2.2) describes the data. Section (2.3) discusses the statistical framework and the identification strategy. Section (2.4) presents estimates of the effects of promotions on heart disease. Section (2.5) discusses and tests potential threats to the validity

of the instrument. Section (2.6) summarizes the results and discusses possible explanations for the observed pattern of effects. Section (2.7) presents estimates of the effects of promotions on other health outcomes. Section (2.8) concludes.

2.2 Data and Descriptive Statistics

The Whitehall II sample contains 10,308 civil servants employed in Inner London from 1985-87. All males and females aged 35 to 55 from 20 Whitehall departments were eligible for inclusion. Overall response rates were approximately 76 percent. The initial data collection consisted of a medical screening and a lengthy questionnaire. Followup data have been collected over four subsequent "phases" since the initial screening, with questionnaires in 1990, 1992, 1996, and 1998, and medical screenings in 1992 and 1998 (each phase takes places over a two to three year period).[2]. Questionnaire data include information on age, gender, employment grade, tenure, marital status, parent and sibling health, education, and self-rated health. Medical measures available for this study include weight, height, the presence of ischemic heart disease, and mortality. Attrition rates in the subsequent phases range from 16 to 24 percent (Marmot and Brunner, 2004). However, for some key measures, such as ischemic heart disease and mortality, there is no attrition because participants' medical records are flagged by the National Health Service (NHS), so medical events are recorded even if participants do not respond to questionnaires.

Employment grades in the British Civil Service were standardized across most departments in 1971 (Her Majesty's Stationery Office, 1971). For research purposes, employment grades were further condensed into six primary grade levels. Ranked from highest to lowest, they are: Unified Grades 1-6 (Administrative), Unified Grade 7 (Administrative), Senior Executive Officer, Higher Executive Officer, Executive Officer, and Clerical/Support Staff.[3] In the existing Whitehall literature, these grades are generally labeled 1 through 6, with 1 being the "highest" grade (Unified Grades 1-6) and 6 being the "lowest" grade (Clerical/Support Staff). In this paper, however, we reverse the numbering so that 1 corresponds to the "lowest" grade and 6 corresponds to the "highest" grade. Although it is inconsistent with the previous Whitehall work, it makes the interpretation of regression coefficients more straightforward.

[2]Two additional phases, in 2001 and 2004, have also occurred. However, I do not have access to these data

[3]Although titles such as "Higher Executive Officer" sound impressive, they in fact refer to relatively low ranking positions.

Table 2.1 reports descriptive statistics. The table is broken into two columns. The first column reports summary statistics for the entire Whitehall II sample. The second column reports summary statistics for individuals that joined the Civil Service during the 1980s and entered at the lowest grade level (Clerical/Support Staff). This subset of civil servants is the focus of much of our analysis; they are superior for analytical purposes because they are least likely to be affected by the sampling frame.[4] In comparing the two samples, females account for 33 percent of the total Whitehall II sample but 64 percent of the primary analytic subsample. This discrepancy occurs because females are concentrated in lower grade positions, and the subsample contains primarily lower grade workers. The subsample also contains less educated workers and lower tenure employees (3.4 years of average tenure versus 17.6 years of average tenure). There is little difference in age between the overall sample and the subsample, however, because both are limited to employees aged 35 to 55. If the overall sample did not select civil servants based on age, the average age of the subsample would surely be lower than that of the overall sample.

2.3 Statistical Framework and Identification Strategy

Two key issues complicate estimation of the effect of promotions on coronary heart disease within the Whitehall II data. First, as in most observational studies, the issue of confounding is a primary concern. Since the "treatment" of interest, promotion, is not randomly assigned, it is likely that promoted and non-promoted individuals differ in important ways that are not caused by the treatment itself. One possibility that has received attention in the literature is "health selection," i.e. the possibility that the causal channels run from health to employment grade because healthy people are more likely to be selected for higher grade positions (Marmot and Davey Smith, 1997; Adda, Chandola, and Marmot, 2003; Chandola, et al., 2003). However, many other possibilities exist. Prior to treatment, promoted individuals may differ from non-promoted individuals in terms of education, family background, psychological disposition, or living environment. All of these factors could independently affect coronary heart disease, confounding interpretation of the results.

The second issue complicating estimation is the sampling frame of the Whitehall II data. In an ideal experiment, each civil servant would enter the dataset as soon as he or she joined the Civil Service and remain in it until time of death. In the Whitehall II data, however, civil servants only

[4]Please see Section (2.3) for further discussion of this issue.

enter the dataset if they are employed in the Civil Service between 1985 to 1987. Because average tenure exceeds 17 years, the sample is selected - individuals are more likely to enter the sample if they remain in the Civil Service for several decades. If promotions positively affect health, this sample selection may attenuate the estimated effect of employment grade on health.[5] On the other hand, failure to receive a promotion could encourage more capable employees to leave the Civil Service if they are not promoted, possibly inducing a bias in the opposite direction.[6] Signing the direction of the overall bias resulting from the sample selection procedure is infeasible.

To address the sampling frame issue, we limit the analysis in most specifications to employees who joined the Civil Service in 1980 or later. These employees have an average tenure of only 3.4 years upon entry to the sample. While it is possible that some sample selection issues remain, their effects should be reduced in comparison to those in the overall sample. We check this conjecture in Section (2.5) using a sample that contains only workers who joined the Civil Service in 1985 or later; it confirms our hypothesis that the 1980+ sample is sufficiently restrictive.[7] We therefore view the 1980+ subsample as a good approximation to the ideal sampling frame that nevertheless maintains a reasonable sample size.

To address the issue of selection into employment grade, we exploit a plausibly exogenous source of variation in employment grade: promotion rates across major Civil Service departments. Since the Northcote-Trevelyan report in 1854, promotion within Civil Service departments has officially been on the basis of merit (Stanley, 2004). However, the distribution of promotion *opportunities* across departments has little relation to merit.[8] The *Civil Service Statistics* state that "vacancies

[5]For example, assume that employment grade is randomly assigned and that employees leave the Civil Service - due to sickness or death - if their health index falls below c. If promotions improve health, then a larger share of the low grade employees will leave the Civil Service prior to the study's start date. These leavers will also tend to be the sicker individuals, so the remaining pool of low grade employees will be drawn from a healthier group of individuals than the remaining high grade employees. The estimated (positive) effect of promotions on health will therefore be attenuated.

[6]Suppose that promotions are randomly assigned and that good employees, whose opportunity cost of employment is greater, leave the Civil Service if they do not receive a promotion within the first six years. Poor employees, in contrast, stay regardless of whether they receive a promotion. The observed pool of promoted employees will therefore consist of a mix of good and poor employees, while the observed pool of non-promoted employees will consist largely of poor employees. If employee quality is positively correlated with health, the estimates could overstate the impact of promotions on health.

[7]The workers joining in 1985 or later have only 1.2 years of tenure on average upon entry to the sample. The results for the 1985+ sample are similar to the results for the 1980+ sample.

[8]The distinction between promotion opportunities and average grade level is key. Average departmental grade level will generally be correlated with employee quality, because departments with more high grade positions will directly recruit high quality employees into those slots. However, conditional on initial grade level, there is no reason to believe that departmental promotion rates are correlated with employee quality. The one exception to this rule are "Fast Stream" employees, who start at lower grade levels with the expectation that they will quickly advance

[within departments] arise through retirements, resignations, promotions to yet higher grades, or through the creation of new posts, offset by any posts that have been lost...There are marked differences between individual departments due to variations in relative grade sizes and in levels of wastage." (HM Treasury, 1985).[9] Furthermore, much of the variation in promotion rates during the period in question arises from the large expansion of hiring that occurred during World War II. This expansion had a differential effect on departments and caused a wave of retirements that occurred from the late 1970s through the mid 1980s (HM Treasury, 1983). Figure 2-1, reproduced from *Civil Service Statistics 1986*, graphically demonstrates the substantial change in the age distribution that occurred during this time period.

Ideally we would use the pattern of wastage at certain grade levels within London as the instrument for departmental promotion rates. However, data of this detail are not available from the Civil Service Statistics Office. Instead we use the observed departmental promotion rate as an instrument for individual promotion.[10] The primary strength of this instrument is that it can be a valid source of identification even if there is selection into departments.[11] Nevertheless, several potential threats to validity exist. We analyze these threats in-depth in Section (2.5); the instrument does not appear to be correlated with other factors that could independently affect health.

2.4 Coronary Heart Disease Results

2.4.1 Cross-Sectional Results

Cross-sectional OLS results for the entire Whitehall II sample show that employment grade is strongly correlated with the presence of coronary heart disease (CHD). Table 2.2 presents results for the regression:

$$CHD_{id} = \beta\ Grade_{id} + X_{id}\delta + \epsilon_{id}$$

The dependent variable is the presence of coronary heart disease, *Grade* is the worker's employ-

through the ranks. We remove the effect of Fast Stream employees by focusing on employees entering at Grade 1 (Fast Stream employees do not enter as clerical/support staff).

[9] "Wastage" refers to worker departures, either through retirement, illness or death, or voluntary departure.

[10] In practice, we implement this instrument by using a set of department dummies as instruments.

[11] The key determinant of validity will be whether the selection into departments is systematically correlated with the future opening of departmental vacancies.

ment grade, and X is a set of controls. Subscript i refers to an individual, and subscript d refers to a department. The first three columns of Table 2.2 report coefficients for the presence of any CHD (CHD that was present upon entering the sample or that occurred between 1985 and 1999) regressed upon grade level. Column (1) controls only for gender, column (2) controls for gender and quadratics in age and potential tenure, and column (3) controls for all covariates in the previous column plus college education (this is the preferred specification). In all three specifications, an increase of one grade level reduces the probability of any CHD by approximately one percentage point. Persons in the highest grade levels are therefore approximately 5 percentage points less likely to have CHD than the lowest grade levels. The coefficients are statistically significant in every specification. Sample size is reduced below 10,308 because not all observations include information on department or college education.[12]

Columns (4) - (6) of Table 2.2 report coefficients for the presence of CHD after the study began (the outcome variable equals one for anyone that had CHD between 1985 and 1999). Again, an increase of one grade level reduces the prevalence of heart disease by about one percentage point. The coefficients in each column are statistically significant.

Table 2.3 presents coefficients for five grade level dummies (the omitted category is the lowest grade level). This specification allows the health gradient to vary across grade levels. The first column presents results for the presence of any CHD; the second column presents results for the presence of CHD after the study began. There appears to be a discrete jump at Grade 3 (Higher Executive Officer), and possibly at Grade 5 (the first Administrative grade). However, the individual grade level dummies do not explain significantly more of the variation in CHD than the single grade level variable (F-statistic of approximately 0.34). We therefore parameterize grade level as a single variable rather than multiple dummies for the remainder of the analysis.

2.4.2 Promotions and Heart Disease

To estimate the relationship between promotions and heart disease, we present specifications that control for grade level at entry into the Civil Service. We also control for the presence of heart disease at entry into the study, so that we are essentially estimating the relationship between changes in grade level and changes in heart disease. However, we do not have information on heart

[12]Running the regressions in the first two columns on the full sample of 10,308 produces similar coefficients.

disease at the exact point of entry into the Civil Service - we are only (approximately) measuring this for workers that entered the Civil Service shortly before the study began. Table 2.4 reports coefficients for the model:

$$CHD_{id} = \beta\ Grade_{id} + \alpha\ CHD\ at\ Entry_{id} + \phi\ Grade\ at\ Entry_{id} + X_{id}\delta + \epsilon_{id}$$

The dependent variable is the presence of coronary heart disease, *CHD at Entry* measures whether the worker had heart disease upon entry to the sample, and *Grade at Entry* is the worker's employment grade upon entry to the Civil Service. The first column in Table 2.4 reports results for the full sample. A promotion of one grade level predicts a 0.3 percentage point reduction in the probability of developing heart disease. This result is not significantly different than zero. However, this insignificance may be due in part to the fact that the CHD measure is being differenced over a much shorter period than the promotion measure. An accurate interpretation of the results is that a promotion of one grade level over a 30 year period does not significantly predict changes in CHD over the last 15 years of that period. When the sample is limited to workers who entered the Civil Service at Grade 1 (Clerical/Support Staff), as reported in Column 2, the magnitude of the coefficient increases to 0.8 percentage points, and the result becomes marginally significant.[13]

The third column in Table 2.4 reports results for employees who entered at Grade 1 between 1980 and 1987. As previously discussed, this sample criterion reduces the selection problem created by the Whitehall II sampling frame. A one grade level promotion now predicts a 3.3 percentage point decline in the probability of developing heart disease. This result is highly significant. The increase in magnitude over previous specifications may be due in part to an attenuation in the sampling frame issue. However, it may also be influenced by the fact that most employees in this sample have not been promoted past Grade 3. The fourth column reports results for employees who entered the Civil Service at Grade 1 prior to 1980, but have not advanced past Grade 3. For this sample, a promotion of one grade level reduces the probability of developing heart disease by 1.8 percentage points, an estimate that is over twice the magnitude of the coefficient for all employees entering at Grade 1.

[13]Like the first column, this column cannot fully control for the prevalence of CHD upon entry into the sample. It is literally reporting that promotion of one grade level over a longer period predicts decreases in CHD over the last 15 years of that period. The coefficient is probably larger than in Column 1 in part because employees who entered at Grade 1 were, on average, more recent hires.

2.4.3 Instrumental Variables Results

We instrument for individual promotions by using the average observed departmental promotion rate for employees who joined the Civil Service at Grade 1 between 1980 and 1987. In practice, we implement this instrument by using a set of department dummies as our instruments. Since we have multiple instruments, we estimate coefficients using two-stage least squares (2SLS). We control for heart disease at entry so that we are effectively instrumenting for changes, as in the previous subsection. The first stage model is:

$$Grade_{id} = Dept_d\gamma + \alpha\ CHD\ at\ Entry_{id} + X_{id}\delta + \nu_{id}$$

Variables are as previously defined, except for *Dept*, which represents a set of department dummies. Because the sample is restricted to employees entering at Grade 1, the dependent variable represents a measure of promotions. First stage results indicate a significant relationship between promotion odds and department assignment. The partial R^2 from a regression of grade level on a set of department dummies is 0.057; the F-statistic on the department dummies is 1.9 ($p = 0.02$). The second stage is:

$$CHD_{id} = \beta\ \widehat{Grade}_{id} + \lambda\ CHD\ at\ Entry_{id} + X_{id}\theta + \epsilon_{id}$$

Variables are as previously defined, except for $\widehat{Grade}$, which is the fitted value for promotions from the first stage. Table 2.5 reports results from the two-stage least squares regression. The coefficients are large and consistent across specifications. They are marginally significant in the first and third columns ($p = 0.08$ and $p = 0.07$ respectively), and statistically significant in the second column ($p = 0.04$). The estimates for the preferred specification (the third column) imply that a promotion of one grade level reduces the probability of heart disease in the subsequent 15 year period by 12.9 percentage points. This value is roughly half the average rate of CHD in departments with the most heart disease and corresponds to an effect size of about 0.40. The magnitude of this coefficient is almost four times larger than the comparable OLS result. Nevertheless, the standard errors are large; we cannot reject the hypothesis that the 2SLS estimator converges to the value of the OLS estimate.[14]

[14]This is slightly different than the hypothesis that the 2SLS and OLS estimators converge to the same value, as

Table 2.6 presents results exploring the sensitivity of the 2SLS regression coefficient to different sample selection criteria. The first column reports results for the primary analytic subsample. The second column reports results from a subsample restricted to employees who joined the Civil Service in 1985 or later. The coefficient is similar in magnitude to the estimate from the primary analytic subsample (-0.136 versus -0.129). The comparability of these two coefficients suggests that the 1980-87 subsample is sufficiently restrictive in addressing the attrition issue discussed in Section (2.3).

The third column in Table 2.6 presents results for a subsample that contains an extra 5-year cohort - it contains all employees that joined the Civil Service at Grade 1 between 1975-87. The coefficient is nearly identical to the estimate from the primary analytic subsample (-0.125 versus -0.129). The fourth column expands the sample to include employees who entered at a higher grade level; it reports results for a subsample that contains employees who entered at Grades 1 or 2 between 1980-87. The coefficient falls in magnitude relative to the primary analytic subsample - from -0.129 to -0.092 - but remains marginally significant ($p = 0.08$). The fifth column reports results for a subsample that contains employees entering at Grades 1 or 2 between 1975-87 - it increases the primary analytic subsample along both the time and grade of entry dimensions. The coefficient is reduced in magnitude by almost one-quarter (-0.099 versus -0.129), but it again remains marginally significant ($p = 0.06$). Finally, the sixth column reports results for a subsample that contains all employees who joined the Civil Service after 1950 in Grades 1-3. The coefficient is -0.049, less than half the magnitude of the estimate from the primary analytic subsample. Nevertheless, the standard error drops to 0.016, so the result is highly significant.

2.5 Potential Threats to Validity

Several potential objections exist to interpretation of the two-stage least squares results as causal effects of promotions on coronary heart disease. All of these issues focus on the possibility of a positive correlation between department promotion rates and employee quality. First, some employees may select into departments based on future expectations of department promotion rates. Second, it is possible that departments directly affect health in ways other than through

the OLS estimator has a non-zero variance and the estimators are positively correlated. A Hausman test does not reject the hypothesis that both estimators converge to the same value. However, the underlying assumption that OLS is efficient may be violated if the data are clustered.

promotions. Finally, the large number of instruments (fifteen) raises concerns about finite sample bias. We analyze all of these issues and conclude that that the results do not appear to be driven by any of them.

2.5.1 Individual Selection

If individuals at Grade 1 select into their departments based upon expectations of future promotion odds, a significant relationship between employee quality and departmental promotion rates may arise. To test whether this possibility is driving the two-stage least squares coefficients, we examine characteristics that are correlated with health but should not be affected by promotions. The results indicate that there does not appear to be any systematic relationship between these characteristics and department choice.

Worker selection into departments affects our estimation strategy only if it is correlated with department promotion rates. Our tests therefore examine the correlation between the average departmental promotion rate (i.e., the instrument) and pre-treatment health characteristics.[15] To implement these tests, we place each outcome on the left-hand side of our two-stage least squares regression and test whether the coefficient on grade level is significant.

To maximize statistical power and correct for multiple inference, we also perform a summary index test that combines all of the pre-treatment health outcomes into a single measure (O'Brien, 1984). If there is a systematic relationship between pre-treatment outcomes and departmental promotion rates, it is more likely to be detected in the summary index. To create the summary index, we demean each outcome and convert it to an effect size by dividing by its standard deviation. This conversion normalizes outcomes to be on a comparable scale. We also switch signs where necessary so that the positive direction always connotes a worse health outcome. We then create a new variable that is the mean of the normalized, demeaned outcomes, and place this variable on the left-hand side of our two-stage least squares regression.

We analyze three broad sets of outcomes: parental conditions, sibling conditions, and individual health characteristics that should not be immediately affected by promotions. For parental and sibling conditions, we estimate results for two different samples. The first sample, reported in the

[15]We use the term "pre-treatment" in this context to refer to outcomes which should not be affected by grade level. We also examine whether there is any relationship between the department dummies and the pre-treatment health characteristics. Overall, we do not find evidence of a significant relationship, but this test lacks statistical power. In the paper, we therefore focus on the relationship between the instrument and pre-treatment health characteristics.

first two columns of Table 2.7, is the normal estimation sample: civil servants entering at Grade 1 between 1980-87. The second sample, reported in the last two columns of Table 2.7, is an expanded estimation sample: civil servants entering at Grades 1-3 between 1950-87. This sample corresponds to the sample used in column 6 of Table 2.6. Although it may be affected by sampling frame issues, we use the larger sample to increase precision and improve the power of our falsification tests. For individual health characteristics we estimate results using only the 1980-87 sample. We cannot use the 1950-87 sample because we have no "pre-treatment" data for civil servants entering in the earlier decades.

The first set of rows in Table 2.7 reports results for parental conditions. The conditions include angina, heart attack, high blood pressure, and diabetes. The results demonstrate that there is no significant relationship between departmental promotion rates and any parental condition. The t-statistic is insignificant for every coefficient in the 1980+ sample. The t-statistics in the 1950+ sample are insignificant as well, except for the coefficient on parental blood pressure, which is marginally significant. However, its coefficient is positive rather than negative. The absence of a negative, significant parental blood pressure coefficient is notable, since parental and sibling blood pressures are the strongest predictors of own heart disease within this set of variables.

The second set of rows in Table 2.7 reports results for sibling conditions. The conditions include angina, heart attack, stroke, high blood pressure, and diabetes. The 1980+ results suggest that there could be a significant relationship between departmental promotion rates and sibling angina - the sibling angina t-statistic just attains statistical significance ($p = 0.05$). However, no other 1980+ coefficient approaches marginal significance; in particular, the sibling blood pressure coefficient is positive and insignificant. The 1950+ coefficients are all insignificant except for the coefficient on sibling heart attacks, which is statistically significant. However, the coefficient is positive, signaling that, if anything, our 2SLS estimates appear to be overly conservative.[16] Furthermore, with a total of twenty-three outcomes tested, we should expect to see at least one or two significant p-values, even if no underlying relationship exists. The summary index test in the last row takes this multiple inference problem into account.

The third set of rows in Table 2.7 reports results for individual health characteristics. The characteristics include self-reported health, a history of heart trouble, height, weight, and the

[16]This significant result may also be due to the sampling frame issues discussed in Section 2.3.

presence of a chronic illness.[17] The results confirm that there is no significant relationship between departmental promotion rates and any health characteristic. No coefficient approaches significance, and several of the results suggest a negative relationship between departmental promotion rates and pre-treatment health characteristics.

The final row in Table 2.7 implements a summary index test that combines all fourteen (or nine) outcomes into a single measure. All outcomes are normalized so that the positive direction connotes a bad outcome. The 1980+ coefficient implies that a one unit increase in the departmental promotion rate improves the overall index by a statistically insignificant 0.06 standard deviations ($t = 0.34$). Although the standard error is large (0.18), a 95 percent confidence interval would barely cover the CHD-employment grade effect size of -0.40 estimated in the previous section. The 1950+ coefficient implies that a one unit increase in departmental promotion rate worsens the overall index by a statistically insignificant 0.02 standard deviations ($t = 0.45$). In comparison, the CHD-employment grade effect size for the 1950+ sample is approximately -0.15.

To further test the robustness of our two-stage least squares results, we examine how the estimated 2SLS coefficient evolves over time. If the effects of a promotion (or denial of promotion) are cumulative over time, then we should expect the estimated effect to be stronger five to ten years after the promotion than one to two years after the promotion. Figures 2-2 and 2-3 plot how the 2SLS coefficient evolves over time for two variables for which we have consistent, repeated measures over time.[18] Figure 2-2 plots the coefficient for self-rated health in 1986, 1990, and 1992 (a higher rating implies better health).[19] As expected, the coefficient increases over time. Figure 2-2 plots the coefficient for self-reported chest pain in 1986, 1990, 1992, and 1996.[20] As expected, the coefficient decreases over time, although the decrease is not monotonic.

Overall, the results suggest that individuals at Grade 1 do not select into departments based on expectations of future promotion odds. There is no consistent relationship between departmental promotion rates and pre-treatment health characteristics, and only one of the fourteen pre-

[17]These variables are measured at entry into the sample. Since employees in our sample joined the Civil Service several years before entry into the sample, these characteristics are not truly "pre-treatment." We therefore focus on characteristics that are reasonably immutable, such as height, or that generally take a substantial amount of time to develop, such as chronic illness. The self-reported health variable is coded from 1 to 5 such that 1 corresponds to excellent health and 5 corresponds to poor health. However, when reporting results I reverse the sign of the coefficient so that a higher score on self-reported health indicates better health.

[18]In every year, we limit the sample to observations that have non-missing data for all years.

[19]The question's definition changes in 1996 and 1998, so the results in those years, not shown, may not be comparable. Nevertheless, the magnitude in 1996 is similar to the magnitude in 1992.

[20]This measure is missing for the last period, 1998.

treatment health outcomes tested is statistically significant. Furthermore, the effects for variables for which we have consistent, repeated measures indicate that the estimated effect of promotions appears to increase over time. This pattern is consistent with a causal effect of promotions on health, and inconsistent with individual selection into departments driving the main results.[21]

2.5.2 Department Effects on Health

Individuals entering high promotion rate departments do not appear less healthy than individuals entering low promotion rate departments. Nevertheless, departments themselves may affect heart disease through other channels than promotion odds. We examine this possibility by testing whether departmental promotion odds capture the entire relationship between department assignment and health. Specifically, we estimate two models with heart disease as the dependent variable. The first model regresses heart disease on a full set of department dummies and other covariates. The second regresses heart disease on the departmental promotion rate and other covariates. We test whether the the unrestricted model (the first) explains significantly more of the variation in heart disease than the restricted model (the second).[22]

An F-test comparing the two models generates a test statistic of approximately 0.39. This result is far below the critical value of 1.75 for this test. We therefore cannot reject the hypothesis that the full set of department dummies explains no more of the variation in heart disease than the departmental promotion rate. This indicates that, at the very least, any effect that departments have on heart disease is highly correlated with departmental promotion rates. In our view, the most likely channel for this effect is through the promotion rate itself. Furthermore, even if departments affected health in other ways than through promotion rates, we would still conclude that the work environment has a significant effect on health, but we would be misspecifying the channels through which the effect was operating.

[21] Another possible explanation for this pattern is that gaps in health status are initially small and grow larger as workers age and general health deteriorates. However, the evidence does not strongly favor this hypothesis. The initial correlation between departmental promotion rates and health is negative, not positive. Furthermore, while the standard deviation in the chest pain variable increases by 38 percent from 1986 to 1996, the standard deviation in the self-reported health variable increases by only 10 percent from 1986 to 1992. Specifying the self-reported health regression in logs rather than levels gives the same pattern of results - the proportional effect is increasing along with the raw effect.

[22] This test is motivated by Autor and Houseman (2005).

2.5.3 Finite Sample Bias

Ideally, we would use promotion rates for entire departments as our instrument. If the entire department were observed, then the sole determinant of departmental promotion rates would be the number of available promotion slots.[23] However, these data do not exist because Whitehall II only sampled employees aged 35 to 55; we observe promotion rates for only a subset of each department. The observed promotion rate is thus a function of both the departmental promotion rate (plausibly exogenous) and the quality of the observed cohorts relative to the unobserved cohorts (assumed endogenous). The latter factor raises concerns about finite sample bias, particularly since we use a relatively large number of instruments (fifteen). If finite sample bias affects the results, it will bias them toward the p-lim of the OLS coefficient (Bound, Jaeger, and Baker, 1995).

To address the finite sample bias issue, we estimate the effect of promotion rates on coronary heart disease using two alternative instrumental variables estimators, jackknife instrumental variables (JIVE) and unbiased split-sample instrumental variables (USSIV). Angrist, Imbens, and Krueger (1999) and Angrist and Krueger (1995) respectively show that these estimators are less affected by finite sample bias. JIVE uses a "leave-one-out" approach when computing the first stage - the first stage for each individual is estimated on a sample that omits that individual's observation. The USSIV estimator uses data from the previous cohort (workers joining from 1975-79) to estimate average departmental promotion rates for the 1980-87 analytic sample.[24] In both cases, the potential correlation between an individual employee's quality and the estimated departmental promotion rate is broken.

Table 2.8 reports results for each of the alternative IV estimators. Column 1 presents results from the original two-stage least squares estimator. Column 2 presents results using JIVE. The estimated coefficient is slightly reduced in magnitude, from -0.129 to -0.109, but it remains marginally significant. Column 3 presents results using USSIV. The coefficient is imprecisely estimated, but it increases in magnitude by 38 percent, from -0.129 to -0.178. Overall, there is no evidence that finite sample bias seriously affects the two-stage least squares results.[25]

[23]This would be true even in finite samples, giving us unbiased two-stage least squares estimates. Our first stage coefficients would be estimated exactly.

[24]Angrist and Krueger (1995) present USSIV in the context of randomly splitting a given sample. In our case, the sample is not randomly split, so an analogy to two sample instrumental variables may seem more appropriate. However, in practice the estimator we implement is the USSIV estimator.

[25]A related source of potential bias in the two-stage least squares estimates is the possibility that departments lobby for extra promotion slots when they have exceptionally skilled cohorts. If so, a systematic correlation between

2.6 Discussion

OLS results suggest that a promotion from Grade 1 (clerical/support staff) to a higher grade will reduce the probability of heart disease by 3.3 percentage points over a subsequent 15 year period. Two-stage least squares results, in comparison, imply a larger reduction of 12.9 percentage points. A number of objections to interpretation of the 2SLS estimates as causal effects exist, including selection into departments, independent effects of departments on health, and finite sample bias. However, we demonstrate that these biases do not appear to be driving the results.

The fact that the 2SLS estimates are of greater magnitude than the OLS estimates may be surprising; many economists would expect that selection effects would cause OLS to overstate the true causal effect. However, we should note that the 2SLS and OLS coefficients do not differ within precision. A true effect located near the OLS coefficient could plausibly generate the observed data even with no bias in either estimator. Nevertheless, there are several reasons why the 2SLS coefficient could converge to a larger value than the OLS coefficient.

One possibility is the presence of measurement error in grade level. Employment grade is measured at only a few points in time; we generally use grade level measured 3 to 8 years after entry into the Civil Service.[26] However, it is presumably the entire history of grade level, rather than grade level at a single point in time, that determines the likelihood of coronary heart disease. Our independent variable of interest is therefore measured with error, which can cause attenuation bias in the context of linear regression. Solon (1992) and Zimmerman (1992), and more recently Mazumder (2005), demonstrate that this type of measurement error generates substantial attenuation bias when estimating the intergenerational elasticity of income.[27] The same phenomenon may be attenuating the OLS estimates in the Whitehall II data. The 2SLS estimator, in contrast, remains consistent in the presence of classical measurement error and therefore will typically be of larger magnitude. Alternatively, if the entire history of promotions is condensed into a single

cohort skill and departmental promotion rates could arise. If this were true, we would expect the USSIV estimates to be smaller in magnitude than the standard 2SLS estimates. In fact, we observe the opposite.

[26]There is a tradeoff between using grade level at additional points in time and reducing attrition; questionnaire attrition becomes increasingly severe in the later phases of Whitehall II. However, the results are qualitatively similar when we use an average value of grade level centered approximately 10 to 15 years after entry into the Civil Service. Using this variable, the results are estimated on a smaller, non-missing sample.

[27]These authors examine intergenerational income mobility. They assume that an individual's income is a function of his or her parents' lifetime income. When regressing individual income on parental income measured at a single point in time (a noisy measure of parental lifetime income), they find much lower coefficients than when using an average of many years of income or when using a two-stage least squares procedure.

binary variable, Angrist and Imbens (1995) demonstrate that the coefficient estimated by instrumental variables may be overstated, although the sign will be correct.[28] This could cause the 2SLS estimate to exceed the OLS estimate.

The 2SLS estimator may also converge to a larger value than the OLS estimator because it estimates the Local Average Treatment Effect (LATE). Angrist, Imbens, and Rubin (1996) establish that instrumental variables estimates the average treatment effect for "compliers," i.e. individuals whose treatment status is affected by the instrument. In the context of our Whitehall instrument, the compliers are employees who are good enough to merit promotion in high promotion rate departments but not in low promotion rate departments.[29] Because these workers know that they are skilled enough to be promoted in a higher promotion rate department, those that are denied promotions may become particularly frustrated, causing deleterious health effects. In fact, these workers may define their social status reference group as the set of other workers of similar skill, rather than other workers in the same department. The results of Deaton and Paxson (1999) therefore imply that the 2SLS estimator, which focuses on workers of similar skill (the relevant reference group), should be less attenuated than the OLS estimator.

Compelling empirical evidence exists in support of this hypothesis. For example, Redelmeier and Singh (2001) demonstrate that Oscar winners live 3.9 years longer on average than Oscar nominees, with a large portion of reduction attributable to a lower rate of ischemic heart disease. This difference, which does not appear to be due to selection or reverse causality, is sizable; it is equal to the difference in life expectancy between the average American and the average Sri Lankan, for example (US Census Bureau, 2005).[30] Furthermore, no similar health gradient appears between Oscar nominees and other actors never nominated for an Oscar. These results therefore support the idea that, for awards and promotions, individuals form their reference groups based on skill. The Oscar nominees who do not win experience a negative shock relative to their reference group, so a research design that compares nominees to other successful actors (instead of to winners) incorrectly codes their social status as above average. If a similar mechanism is at work for Whitehall employees,

[28]The promotion variable is not literally coded as binary. Nevertheless, it only take on a small number of discrete values, and the true effect may work though an entire history of promotions.

[29]Employees who are not compliers fall into the categories of "always-takers" and "never-takers." The always-takers are employees who are skilled enough to be promoted even in low promotion rate departments. The never-takers are employees who are not skilled enough to be promoted even in high promotion rate departments.

[30]Miskie, Near, and Hegele (2003) report similar results for Nobel prize winning scientists, though these findings are more likely to be affected by reverse causality, as the Nobel is awarded towards the end of a scientist's life span.

we should expect the 2SLS estimates to exceed the OLS estimates.

It is also possible that OLS is upwardly biased (biased towards zero) for conventional omitted variables reasons. A common initial assumption is that upper grade employees differ from lower grade employees in ways that improve health status, such as additional education and better family background. However, there is limited empirical evidence to support this selection hypothesis. Controlling for college education in the OLS results, for instance, reduces the magnitude of the grade level coefficient in some specifications and raises it in others. The change in the grade level coefficient is never substantial. Furthermore, while there is surely some positive selection by grade level, negative selection factors are also likely to be present. For example, employees entering the Civil Service at high grades are actually more likely to have a family history of heart attacks or strokes. Negative selection factors such as this one may offset positive selection factors, and the net omitted variables bias could run in either direction.

Finally, it is plausible that promotions could have positive external effects on the coworkers of employees who are promoted. For example, workers in departments with high promotion rates may be happy because they believe that they too will soon receive a promotion. In that case, the OLS estimate would tend to understate the true effect, because some non-promoted employees would receive a beneficial treatment when their coworkers were promoted. The 2SLS estimate, in contrast, would tend to overstate the effect for a treated individual, because it assumes that the entire effect is operating only through individuals that are promoted. The 2SLS coefficient would therefore be estimating the net internal and external effects of a promotion on heart disease, rather than simply an internal effect.

The existing literature on an experimental manipulation of socioeconomic status on health is limited. However, the estimates reported in this paper are consistent, in sign, magnitude, and pattern, with the estimates reported in other studies. The first set of studies measures the effect of social status on health in non-human primates. For example, Shively and Clarkson (1994) manipulated social status within a group of female cynomolgus monkeys. They allowed the monkeys to initially sort into a linear social hierarchy, and then placed the dominant monkeys in one group and the subordinate monkeys in a different group. The two new groups were allowed to re-sort, so that half of each new group became dominant and half subordinate. Shively and Clarkson report that the monkeys that went from dominant to subordinate had 111 percent more coronary artery

plaque than the monkeys that went from subordinate to dominant. This result indicates that the manipulation of social status, rather than an initial selection effect, significantly increased coronary artery plaque in the affected monkeys.

The quasi-experimental research on human primates reaches similar conclusions. In particular, Lleras-Muney (2005) uses changes in compulsory schooling laws to estimate the effect of education on mortality. She concludes that one additional year of education reduces the probability of dying in the next decade by approximately 3.6 percentage points. Like our results, she reports 2SLS estimates that are several times larger than least squares estimates. While it is difficult to make a direct comparison of results because the outcome variables are different, a rough computation suggests that the estimated effects are similar. For example, if the entire effect were assumed to run through the channel of income, the elasticity of heart disease (or, in Lleras-Muney's case, mortality) with respect to income would be around 3 in either case. These elasticities are an order of magnitude larger than the elasticity of mortality with respect to income that Deaton and Paxson (1999) estimate, suggesting that in both cases income may not be the main causal channel.

2.7 Other Health Outcomes

We examine a range of other health outcomes to determine whether promotions have a broader impact on health beyond reducing heart disease. The outcomes we analyze include self-reported health, mortality rates, self-reported heart trouble, SF-36 survey physical and mental health component scores, and marital rates. Promotions are correlated with a beneficial response for physical outcomes, but a negative response for mental outcomes.

Self-reported health is measured on a scale of 1 to 5. In the original coding of the variable, 1 corresponds to excellent health and 5 corresponds to poor health. However, for ease of interpretation we switch the sign so that a higher self-reported health score corresponds to better health. The first and second columns of Table 2.9 report results for OLS and 2SLS regressions respectively. The first column regresses the health outcome at the observation date on grade level, an initial measurement of the health outcome, and the full set of controls used in previous models.[31] As in previous specifications, the sample is restricted to employees who started at the first grade level and

[31]There is no (meaningful) initial measurement of the mortality and the SF-36 variables. In those cases, we control instead of initial self-reported health.

joined the Civil Service in the 1980s. The coefficients reported therefore correspond to promotion effects. The second column reports the two-stage least squares analog of the model in the first column; the instrument is the departmental promotion rate for the 1980-87 cohort. The third column reports the date at which the health outcome was observed.

The results in the first row of Table 2.9 suggest that promotions may improve self-reported health. The OLS coefficient indicates that promotions are associated with positive changes in self-reported health, but the coefficient is insignificant. However, the 2SLS coefficient implies a stronger effect on self-reported health (0.55 versus 0.17), and the effect approaches marginal significance. The results therefore suggest that promotions could improve self-reported health by as much as 0.6 standard deviations, but they are imprecisely estimated.[32]

The second outcome we examine is mortality. Mortality data are collected from medical records tagged by the NHS, so there is virtually no attrition for this variable. The results in the second row of Table 2.9 suggest that promotions could reduce mortality, but the coefficients are not of a consistent sign. The OLS coefficient indicates that promotions are associated with a marginally significant reduction in the probability of death (approximately 1.4 percentage points). The 2SLS coefficient, in contrast, implies that promotions increase the probability of death (the estimated effect is quite large - over 9 percentage points). However, this coefficient is imprecisely estimated; the standard error exceeds the coefficient in magnitude and is so large that no reasonable effect could ever be statistically significant. The evidence is therefore inconclusive.

The third outcome we examine is self-reported heart trouble. This variable is collected from questionnaire data - individuals are asked whether they have ever experienced any heart trouble. It therefore exhibits more attrition than the standard CHD variable. The results in the third row of Table 2.9 demonstrate that promotions reduce self-reported heart trouble. The OLS coefficient shows that promotions are associated with a 2.4 percentage point reduction in self-reported heart trouble; this effect is statistically significant. The 2SLS coefficient implies that promotions reduce self-reported heart trouble by a statistically significant 16.8 percentage points. It is notable that these estimates are roughly similar in magnitude to the coronary heart disease estimates reported in Section (2.4). This fact suggests that the self-reported results, with their higher attrition rate, may be as reliable as the clinically measured heart disease variable.[33]

[32]The self-reported health variable has a standard deviation of approximately 0.93 in the estimation sample.

[33]It also suggests that the somewhat different time scales of the two variables do not qualitatively change the

The fourth and fifth outcomes we examine are the physical and mental component scores of the Short Form 36 survey (SF-36). The SF-36 is a commonly used health survey designed to measure quality-of-life. Sample items include questions like, "Does your health limit you when climbing several flights of stairs?" Higher scores are better. The results in the fourth row of Table 2.9 demonstrate that promotions improve physical health as measured by the physical component of SF-36. The OLS coefficient shows that promotions are associated with a 0.8 point increase in the SF-36 physical health score. The 2SLS coefficient implies that promotions increase the SF-36 physical health score by 3.8 points. This result is statistically significant and corresponds to an effect size of approximately 0.43. However, the fifth row demonstrates that, if anything, promotions negatively impact mental health as measured by the mental component of SF-36. The OLS coefficient shows that promotions are associated with a 1.3 point decrease in the SF-36 mental health score, and the 2SLS coefficient implies that promotions decrease the SF-36 mental health score by 3.9 points (an effect size of about 0.41). The 2SLS estimate approaches marginal significance.[34]

The last outcome we examine is marital rates. This variable is collected from questionnaire data; our variable measures whether an individual reports ever having been married. The results in the last row of Table 2.9 suggest that, if anything, promotions reduce marriage rates. The OLS coefficient indicates that promotions are associated with a reduction in marriage rates (approximately 2 percentage points), but the coefficient is not significant. The 2SLS coefficient implies a stronger, negative relationship between promotions and marriage rates (the magnitude grows to approximately 7.4 percentage points). It approaches marginal significance.

2.8 Conclusion

We use departmental promotion rates as a credibly exogenous source of variation in employment grade to estimate the causal effect of promotions on coronary heart disease. Our estimates are sizable, implying that a promotion from the lowest grade level reduces the prevalence of heart disease by 3.3 to 12.9 percentage points. These estimates do not appear to be driven by employee selection or endogeneity of departmental promotion rates, although our robustness checks are limited in

results. This seems plausible since even the self-reported measure is still measured at least 7 to 10 years after the promotion.

[34]Mental health may be negatively affected by the fact that workers in higher grade positions report their jobs as being more stressful, even though they also report having more job control.

their power. Nevertheless, these effects are reasonably consistent with other estimates of the causal effects of socioeconomic status on health. They also provide empirical evidence in support of the hypothesis that social status is often defined in comparison to a narrow reference group of peers. Even for individuals with differences in social status that appear small on a global scale, a steep health gradient can therefore appear .

Figure 2-1: Changes in the Civil Service Age Distribution from 1978 to 1986

Notes: Reproduced from *Civil Service Statistics 1986.*

Figure 2-2: Effects of Promotion on Self-Reported Health Over Time

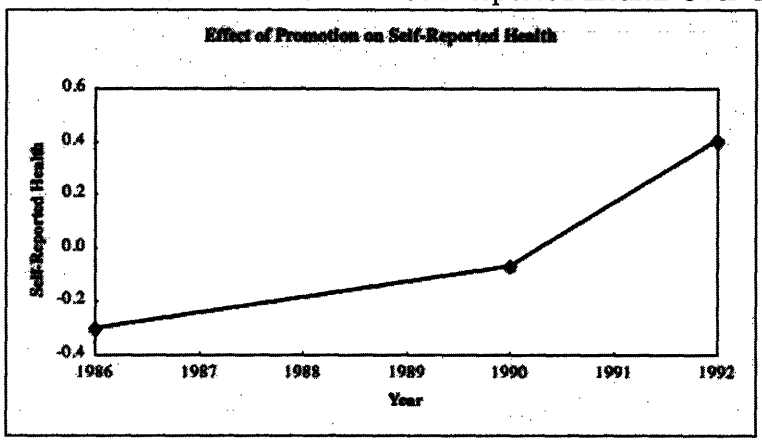

Figure 2-3: Effects of Promotion on Chest Pain Over Time

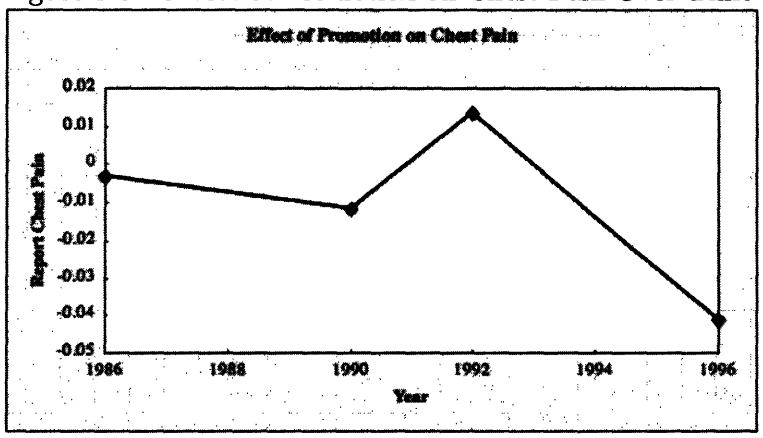

Table 2.1: Summary Statistics

Variable	Full Sample	Subsample
Grade Level (1-6)	3.23	1.16
	(1.69)	(0.50)
Female	0.331	0.642
	(0.471)	(0.480)
Age	44.4	44.8
	(6.1)	(6.2)
Tenure	17.6	3.4
	(8.5)	(2.2)
College	0.469	0.284
	(0.499)	(0.452)
Any CHD	0.133	0.134
	(0.340)	(0.341)
Sample Size	10,308	710

Notes: Parentheses contain standard deviations.

Table 2.2: Cross Sectional Relationship Between Employment Grade and CHD

	Any CHD			CHD Post-1985		
	(1)	(2)	(3)	(4)	(5)	(6)
Grade Level	-0.0107	-0.0115	-0.0105	-0.0091	-0.0098	-0.0096
	(0.0020)	(0.0022)	(0.0026)	(0.0021)	(0.0023)	(0.0026)
R^2	0.003	0.022	0.022	0.002	0.021	0.019
N	8,822	8,670	6,837	8,822	8,670	6,837
Controls:						
Age, Tenure	No	Yes	Yes	No	Yes	Yes
College	No	No	Yes	No	No	Yes

Notes: All models control for gender. Parentheses contain standard errors clustered at the department level.

Table 2.3: Cross Sectional Relationship Between Each Grade and CHD

Grade	Any CHD	Post-1985 CHD
Grade 6	-0.044	-0.042
	(0.016)	(0.016)
Grade 5	-0.044	-0.040
	(0.011)	(0.012)
Grade 4	-0.023	-0.031
	(0.014)	(0.017)
Grade 3	-0.023	-0.026
	(0.017)	(0.016)
Grade 2	0.002	-0.002
	(0.018)	(0.017)
R^2	0.022	0.019
N	6,837	6,837
Controls:		
Age, Tenure	Yes	Yes
College	Yes	Yes

Notes: All models control for gender. Parentheses contain standard errors clustered at the department level.

Table 2.4: Relationship Between Promotions and Changes in CHD

	(1)	(2)	(3)	(4)
Grade Level	-0.0028	-0.0080	-0.0330	-0.0181
	(0.0029)	(0.0041)	(0.0126)	(0.0134)
R^2	0.117	0.108	0.107	0.120
N	6,138	3,564	494	2,443
Sample Restrictions:	None	Enter at Grade 1	Enter at Grade 1 Enter 1980-87	Enter at Grade 1 Achieve Grade 1-3 Enter Before 1980

Notes: All models control for gender, age, tenure, and college education. Parentheses contain standard errors clustered at the department level.

Table 2.5: 2SLS Relationship Between Promotions and CHD

	(1)	(2)	(3)
Grade Level	-0.102	-0.137	-0.129
	(0.054)	(0.061)	(0.067)
R^2	0.098	0.090	0.093
N	494	494	494
Controls:			
Age, Tenure	No	Yes	Yes
College	No	No	Yes

Notes: All models control for gender. Parentheses contain standard errors clustered at the department level.

Table 2.6: 2SLS Relationship Across Different Subsamples

	(1)	(2)	(3)	(4)	(5)	(6)
Grade Level	-0.129	-0.136	-0.125	-0.092	-0.099	-0.049
	(0.067)	(0.091)	(0.062)	(0.050)	(0.050)	(0.016)
R^2	0.093	0.144	0.064	0.117	0.056	0.090
N	494	198	983	576	1,294	5,364
Joined Civil Service:	1980+	1985+	1975+	1980+	1975+	1950+
Grade of Entry:	1	1	1	1-2	1-2	1-3

Notes: All models control for gender, age, tenure, and college education. Parentheses contain standard errors clustered at the department level.

Table 2.7: Falsification Tests - Relationship Between the Instrument and Other Outcomes

	1980+, Entry Grade 1		1950+, Entry Grade 1-3	
	2SLS Coefficient	N	2SLS Coefficient	N
Parental Outcomes				
Angina	-0.025	466	0.008	5,163
	(0.201)		(0.011)	
Heart Attack	-0.198	473	-0.015	5,189
	(0.196)		(0.025)	
High Blood Pressure	0.173	467	0.037	5,195
	(0.217)		(0.020)	
Diabetes	-0.121	466	-0.010	5,127
	(0.114)		(0.018)	
Sibling Outcomes				
Angina	-0.077	409	0.012	4,325
	(0.036)		(0.008)	
Heart Attack	-0.033	412	0.025	4,332
	(0.099)		(0.009)	
Stroke	-0.041	410	-0.008	4,317
	(0.038)		(0.007)	
High Blood Pressure	0.131	411	0.018	4,338
	(0.092)		(0.017)	
Diabetes	-0.040	413	0.011	4,323
	(0.057)		(0.009)	
Own Outcomes				
Self-Reported Health	-0.358	494		
	(0.304)			
Ever Heart Trouble	0.063	491		
	(0.084)			
Height	2.35	494		
	(2.41)			
Weight	-0.70	494		
	(3.84)			
Chronic Illness	0.262	490		
	(0.236)			
Summary Index	-0.059	494	0.017	5,333
	(0.176)		(0.038)	

Notes: All models control for gender, age, tenure, and college education. Parentheses contain standard errors clustered at the department level.

Table 2.8: Exploring Finite Sample Bias - Alternative 2SLS Specifications

	2SLS	JIVE	USSIV
Grade Level	-0.129	-0.109	-0.178
	(0.067)	(0.061)	(0.249)
R^2	0.093	0.107	0.075
N	494	494	494
Controls:			
Age, Tenure	Yes	Yes	Yes
College	Yes	Yes	Yes

Notes: All models control for gender. Parentheses contain standard errors clustered at the department level.

Table 2.9: Relationship Between Promotions and Other Health Outcomes

	OLS	2SLS	Observation Date	N
Self-Reported Health	0.17	0.55	1991-93	365
	(0.12)	(0.34)		
Mortality	-0.014	0.091	1985-99	493
	(0.008)	(0.107)		
Ever Heart Trouble	-0.024	-0.168	1991-93	378
	(0.012)	(0.082)		
SF-36 Physical Health	0.84	3.82	1991-93	363
	(0.75)	(1.80)		
SF-36 Mental Health	-1.30	-3.94	1991-93	363
	(1.17)	(2.35)		
Ever Married	-0.020	-0.074	1995-96	317
	(0.015)	(0.044)		

Notes: All models control for gender, age, tenure, and college education. Parentheses contain standard errors clustered at the department level.

Chapter 3

Safety For Whom? The Effects of Light Trucks on Traffic Fatalities

3.1 Introduction

Over the course of two decades light truck sales have increased from 21.8 percent of light vehicles sales in 1980 to 56.6 percent of light vehicles sales in 2004 (Ward's Reports Incorporated, 1980-2005). This sales growth has induced a corresponding shift in vehicle fleet composition that continues today. The commercial success of these vehicles is due in part to the fact that many consumers associate them with occupant safety because of their high mass and rugged image (Bradsher, 2002). Indeed, previous research has suggested that a heavier vehicle fleet is a safer vehicle fleet (Crandall and Graham, 1989). Nevertheless, the net effect of light trucks on traffic fatalities - the leading cause of death for persons under the age of 40 - is ambiguous.

It is well-established in the highway safety literature that additional vehicle mass can improve own-occupant outcomes in the event of an accident, all other things being equal (Evans and Frick, 1993; Evans and Frick, 1994). However, in multi-vehicle collisions, the safety benefits of additional mass come at the cost of greater external risk to the occupants of other vehicles. The net effect of additional vehicle mass on traffic fatalities can therefore be minimal (Evans, 2001).[1] Furthermore,

[1]This fact explains how the 2,600 pound Honda Civic can achieve a 5-star government crash rating while the 6,800 pound GMC Yukon only achieves a 4-star rating (National Highway Traffic Safety Administration, 2006). Since the government test simulates a head-on collision with a vehicle of similar make and model, these experimental results suggest that a vehicle fleet composed entirely of Civics could be safer than a vehicle fleet composed entirely of Yukons. This is possible because the loss in own-occupant protection in the lighter Civics is offset by a corresponding decrease

light trucks possess several unique features that may make them more dangerous both to their own occupants and to others. Relatively poor braking and maneuverability make them difficult to handle, and a high center of gravity increases the risk of rollover. Stiff frames and high ground clearance make them more likely to ride up over car bumpers and penetrate passenger compartments. Because of these distinct characteristics, and because the increased popularity of light trucks has coincided with a deceleration in traffic fatalities, the impact of light trucks on traffic safety has attracted a sizable body of research.

To measure the effects of light trucks on traffic fatalities, I first estimate the net effect of vehicle fleet composition on traffic fatalities using a state-level panel data set spanning 24 years. I then use a random sample of police-reported accidents to estimate the effects of light trucks on their own occupants and on other roadway users when an accident occurs. The risks that light trucks impose upon their own users I define as "internal" risks; the risks that light trucks impose upon other vehicles and pedestrians I define as "external" risks.[2] This distinction is important for policy - rational consumers will account for increased internal risk when making their purchasing decisions, but they will not account for increased external risk unless corrective taxes are in place.[3] I combine the results from the state-level and accident-level estimation strategies to determine an important unknown parameter: the relative crash rate of light trucks as compared to cars. Using this parameter, I estimate the total internal and external effects of shifting the vehicle fleet composition from cars towards light trucks.

Previous analyses using accident-level micro data have shown that, in the event of an accident, light trucks pose a greater hazard to other vehicles than do cars (Joksch, 1998; Gayer, 2004; White, 2004). In particular, White (2004) uses a random sample of all police-reported accidents to demonstrate that when an accident occurs, light trucks impose significant externalities on other cars, trucks, and pedestrians. However, a vehicle can be hazardous for two reasons: it can inflict more damage when a crash occurs and it can be more likely to have a crash. Accident-level data alone cannot identify the relative crash rate of light trucks as compared to cars.[4] Without a reasonable estimate of this parameter, it is impossible to estimate the total effects of light trucks

in the risk that the vehicles impose upon other drivers.

[2]Note that, because collisions involving two light trucks are not uncommon, some of the external risks are borne by the occupants of other light trucks.

[3]In practice, it is unclear whether consumers accurately estimate the internal safety benefits and risks of light trucks.

[4]Previous research, such as Evans (1985), has suggested that larger vehicles are more likely to crash.

on traffic fatalities.[5]

An alternative approach is to use county- or state-level data to estimate the effect of vehicle fleet composition on total traffic fatalities. Coate and VanderHoff (2001) use several years of state-level data to measure the effect of light trucks on traffic fatalities and find inconsistent results across different specifications.[6] However, with state-level data alone it is impossible to disaggregate the net fatality effect into its internal and external components.

This paper is organized as follows. Section (3.2) describes the data. Section (3.3) presents an empirical model and results for the net effect of vehicle fleet composition on traffic fatalities using state-level panel data. Section (3.4) presents an empirical model and results for the internal and external effects of light trucks on fatalities. Section (3.5) concludes. The results suggest that a one percentage point increase in the light truck share of the vehicle fleet increases annual traffic fatalities by approximately 172. This estimate appears robust to a variety of concerns, including vehicle fleet endogeneity and unobserved state trends. Of this increase, approximately 66 to 75 percent are external fatalities, and the remaining 25 to 34 percent are internal fatalities.

3.2 Data and Descriptive Statistics

I use two distinct data sets to estimate the effects of light trucks on traffic fatalities. The first data set is a state-level panel data set from the Federal Highway Administration's Highway Statistics Series (U.S. Department of Transportation, 2004). The data include information on traffic fatalities, vehicle registrations, vehicle miles traveled, and licensed drivers.[7] I supplement this information with demographic, state policy, and weather data from the U.S. Census, the Current Population Survey, the National Highway Traffic Safety Administration, and the National Oceanic and Atmospheric Administration. The data span from 1981 to 2004.

The second data set is a random sample of police-reported accidents. These data come from

[5]White (2004) presents a wide range of possible internal, external, and total effects using different values of the relative crash rate parameters. Gayer (2004) has data on fatal accidents and uses a combination of pedestrian fatalities and snow depth to estimate the relative crash rate parameters. The study's identification exploits the assumption that, conditional on an accident occurring, the probability of a pedestrian fatality is uncorrelated with vehicle type. However, White (2004) appears to refute this assumption.

[6]While analyzing the effects of various policies on highway fatalities using of county-level data in 1970 and 1980, Keeler (1994) includes light trucks as a control variable. There is some evidence that light trucks may increase fatalities.

[7]Traffic fatalities include driver, passenger, motorcyclist, pedalcyclist, and pedestrian deaths.

the National Highway Traffic Safety Administration's General Estimates System (GES).[8] The data include information on injuries and fatalities, vehicle type, geographic location, weather conditions, use of safety equipment, and driver and occupant characteristics. For analytic purposes, I decompose this data set into five sub-samples: two-vehicle crashes involving at least one car, two-vehicle crashes involving at least one light truck, single-vehicle crashes, crashes involving motorcycles, and crashes involving pedestrians or pedalcyclists. These data span from 1996 to 2004.

Table 3.1 presents summary statistics for the two data sets. There was a mean of 881 annual fatalities per state during the sample period, and the light truck share of the vehicle fleet averaged close to 30 percent.[9] 47 continental states are included in the analysis; Oklahoma is excluded because it is missing data for some years. The accident-level data set includes data on 454,014 accidents.[10] Of these accidents, 5,717 involve at least one fatality, and approximately 47 percent of them involve at least one light truck.

3.3 State-Level Results

3.3.1 Results

The popularization of the sport utility vehicle has resulted in strong growth in light truck sales over the past 15 years. After recovering from the oil price shocks of the early 1980s, light truck sales accounted for 30.2 percent of new vehicle sales in 1985. Market share gains continued until 1991, when the light truck share of sales reached 34.8 percent. Growth then accelerated, and by 2004 light trucks constituted 56.6 percent of new vehicle sales (Ward's Reports Incorporated, 1980-2005). This rapid growth in market share provides an opportunity to examine the relationship between traffic fatalities and light truck share across different states over time.

The number of traffic fatalities is specified as a function of the light truck share of a state's vehicle fleet and a set of observed and unobserved variables:

$$E[F_{st}|LTS_{st}, POP_{st}, X_{st}, \gamma_t, \alpha_s] = exp[\beta\ LTS_{st} + ln(POP_{st}) + X_{st}\delta + \gamma_t + \alpha_s] \qquad (3.1)$$

[8]This is the same data source used in White (2004), but with additional years.

[9]These averages are unweighted, so they do not correspond to national averages.

[10]Not all of these accidents will necessarily make it into one or more of the analytic data sets. For example, a single-vehicle accident involving a heavy truck would not be present in any of the analytic data sets.

F_{st} represents the number of traffic fatalities in state s during year t, LTS_{st} is the percentage of vehicle registrations that are light trucks, POP_{st} is state population (the scale variable), and γ_t and α_s are unobserved year and state effects respectively. The vector X_{st} includes per capita beer consumption, indicators for primary and secondary safety belt laws, share of vehicle miles traveled that are on rural roadways, percent of licensed drivers under 25, percent of licensed drivers over 70, rural and urban speed limits, the unemployment rate, share of the population with a college degree, share of the population under 18, and measures of rainfall and snowfall.

Because the dependent variable takes on integer values, I use a negative binomial regression model for estimation purposes.[11] To account for unobserved state and year effects, I include indicators for both states and years in all regressions.[12] The strength of the negative binomial model is that it is unbiased if the conditional mean is correctly specified. In that sense, it is no less robust than least squares, and it may provide a better approximation of the conditional expectation function. In theory, the standard errors could still be sensitive to the maximum likelihood assumptions, but experimentation with bootstrapped standard errors produced estimates of similar size to the analytic standard errors. Nevertheless, I also present results for linear regressions to verify that the conclusions are not overly sensitive to the estimation procedure.

Columns (1) through (4) of Table 3.2 present results for the negative binomial models. The regressions are weighted by state population, and the standard errors are clustered at the state-level to correct for within-state serial correlation (Bertrand, Duflo, and Mullainathan, 2004). Column (1) reports results from the preferred specification, which contains a full set of controls. The coefficient of 0.414 on the light truck share variable is statistically significant and implies that a one percentage point increase in light truck share is associated with a 0.41 percent increase in traffic fatalities. The scale variable - population - has a coefficient close to one (1.03), as expected. Although not reported in the table, other control variables have coefficients with signs in the expected directions. Beer consumption has a positive and significant coefficient, while primary belt laws, the unemployment

[11] Michener and Tighe (1992) estimate a Poisson regression model of highway fatalities and find evidence of overdispersion. I therefore employ a negative binomial model; see Hausman, Hall, and Griliches (1984) for details. In theory, fatal accidents is a more appropriate dependent variable because a small percentage of accidents include more than one fatality. However, the Highway Statistics Series stopped publishing the fatal crash series in 1998. Running the primary specification on pre-1998 data, the coefficient changes by less than 2 percent when using fatal accidents instead of fatalities as the dependent variable. For the purposes of this regression, it is therefore safe to use the two measures interchangeably.

[12] As noted in Allison and Waterman (2002), the use of dummy variables in negative binomial models does not appear to be affected by the incidental parameters problem.

rate, and the share of population under 18 have negative and significant coefficients.[13]

Column (2) reports results from a specification containing a base set of controls: population, beer consumption, safety belt laws, unemployment, rural mileage, and weather. The coefficient on light truck share (0.362) remains positive and statistically significant. Column (3) reports results from the preferred specification when the sample is limited from 1996 to 2004. This specification is of interest because the accident-level data set covers the years from 1996 to 2004.[14] The coefficient on light truck share (0.415) is nearly identical to the coefficient in column (1) and attains marginal significance ($z = 1.76$). Column (4) reports results from a specification that controls only for population, year effects, and state effects. The light truck share coefficient of 0.420 is very close to the coefficient in the preferred specification, but the standard error has increased to the point at which the result is no longer statistically significant.

Columns (5) and (6) in Table 3.2 present weighted linear regression models that are directly analogous to the negative binomial models in columns (1) and (2). As in columns (1) and (2), the light truck share coefficients in columns (5) and (6) are positive and significant, although they have decreased in magnitude by 17 percent.[15] Nevertheless, the linear regressions generate estimates that are similar to the negative binomial regressions, but slightly less precise.

A causal interpretation of the estimate from the preferred specification, 0.414, implies that that the 21.8 percentage point increase in light truck share from 1981 to 2004 has increased traffic fatalities by approximately 3,500 deaths in 2004. However, there are several reasons to doubt that the coefficient reported in column (1) represents an estimate of the causal effect of light trucks on traffic fatalities. In particular, issues including the endogeneity of light trucks purchases, measurement error in the light truck series, changes in vehicle miles traveled, and unobserved changes in traffic fatality rates may affect the coefficient estimate.

[13]In the cross-section, one might expect the unemployment rate to be a proxy for economic development, resulting in a positive association with traffic fatality rates. However, since all regressions contain state fixed effects, the unemployment rate is instead a proxy for local recessions, which one might expect to be negatively correlated with traffic fatality rates. This relationship is consistent with previous research, including Evans and Graham (1988), Ruhm (1995), and Ruhm (2000).

[14]In addition, the Federal Highway Administration's definition of a light truck changed somewhat prior to 1996. The results in column (3) suggest that this definitional change is not affecting the estimated coefficient.

[15]Technically the coefficient in column (4) is right at the margin of statistical significance, with a t-statistic of 1.95.

3.3.2 Potential Issues in Causal Interpretation

Vehicle Fleet Endogeneity

Many consumers equate light trucks with superior occupant protection. It is therefore possible that light truck sales are partially driven by the fatality rate itself; consumers may switch to vehicles that they perceive to be safer when the accident rate rises. I explore this possibility by examining the relationship between traffic fatalities and new vehicle sales.

Using sales data, I check whether light truck sales respond to changes in the state traffic fatality rate. Columns (1) through (3) of Table 3.3 present the results from regressions of the light truck share of vehicle sales on contemporaneous and lagged values of traffic fatalities per capita, as well as a full set of controls.[16] Column (1) specifies the truck share variable in levels and the fatality variables in logs. Column (2) specifies both variables in logs, and column (3) specifies both variables in levels. Column (4) estimates the preferred traffic fatalities specification, equation (3.1), for the same time period as columns (1) through (3).[17] If the positive coefficient on light truck share in the original regressions is due to the endogeneity of light truck sales, then the regressions in columns (1) through (3) should show an even stronger relationship between light truck sales and fatalities per capita than the regression in column (4). However, I find the opposite pattern. None of the fatality variable coefficients is statistically significant, and a test of the hypothesis that the sum of the fatality coefficients equals zero fails to reject in all three columns, producing p-values of 0.49, 0.98, and 0.15 respectively. In comparison, the light truck share coefficient in column (4) has a p-value of 0.02. It is therefore unlikely that the positive coefficient on light truck share is due to the endogeneity of light truck sales.

Measurement Error and Changes in Vehicle Miles Traveled

Measurement error is present in the light truck registration series, which is based on data reported by state authorities. Because not all states register vehicles on the same time frame, the Federal Highway Administration adjusts the series to make them comparable across states. These adjust-

[16] Note that if causality runs from light truck share to fatalities, then lagged fatality values should not be related to current light truck sales except through serial correlation.

[17] To facilitate comparison, I estimate the coefficient in column (4) using least squares. Columns (1) through (3) have fewer observations than the regressions in Table 3.2 for two reasons. First, the lagged fatality values eliminate several years at the beginning of the sample. Second, Ward's Auto stopped reporting data on registrations of new vehicles in 2001.

ments are imperfect, introducing a degree of error into the series. In general this measurement error should attenuate the coefficient on the light truck variable, biasing it towards zero.[18] The direction of the bias therefore implies that, if anything, the estimated coefficient on the light truck share variable is too conservative.

Of greater concern is the possibility that vehicle miles traveled within a state may be positively correlated with a state's light truck share. To explore this possibility, I include the log of total vehicle miles traveled as a scale variable in addition to population. The results from this specification are presented in column (2) of Table 3.4. The coefficient on light truck share is very close to the original preferred specification, displayed in column (1). It is slightly reduced in magnitude, from 0.414 to 0.401, but it remains positive and statistically significant ($z = 2.45$).[19] Furthermore, a regression of log vehicle miles traveled on light truck share and a full set of controls produces a coefficient of 0.028 with a t-statistic of 0.18.[20] I therefore find no evidence of a relationship between vehicle miles traveled and light truck share.

Unobserved Changes in Traffic Fatality Rates

All models include state effects to control for unobserved differences across states that are constant over time. Nevertheless, unobserved factors that affect traffic fatalities and change over time may be correlated with light truck share growth. Specifically, traffic fatalities depend upon three factors: roadway conditions, vehicle technology, and driver characteristics.[21] Any unobserved factor that affects the coefficient on light truck share should therefore operate through one of these three channels.

It is possible that states which had greater levels of growth in their light truck share experienced less favorable demographic changes or road improvements than states with lower levels of growth in their light truck share. To check for this possibility, I perform several tests. First, I include a

[18]The bias could be exacerbated by the inclusion of state effects, depending on the degree of serial correlation within the measurement error.

[19]It is possible that including vehicle miles traveled may downwardly bias the light truck coefficient. The vehicle miles traveled series is primarily based off of state gasoline sales. Although the government makes some adjustments for vehicle fleet composition, they may be insufficient to account for the substantially lower mileage of light trucks as compared to cars. Therefore, the government may overestimate vehicle miles traveled in states with high growth in light truck share. Such an overestimate would downwardly bias the light truck share coefficient.

[20]This estimate is economically insignificant as well as statistically insignificant. The point estimate implies that a 10 percentage point change in light truck share increases vehicle miles traveled by only 0.28 percent.

[21]Medical technology could also play a role, but for purposes of this analysis it can be subsumed under driver characteristics.

set of state time trends in the regression. The results from this regression are presented in column (3) of Table 3.4. Comparing column (1) of Table 3.4, which has no state time trends, with column (3), which includes state linear time trends, reveals that the light truck share coefficient increases by 20 percent and remains statistically significant. This suggests that geographic trends that are correlated with light truck purchases are not upwardly biasing the coefficient on light truck share.

Nevertheless, the technology of a state's vehicle fleet could plausibly be correlated with light truck share in a manner that would not be accounted for by time trends. In particular, because the growth in light truck share is a recent phenomena, states which have a higher rate of turnover in their vehicle fleets will have more light trucks. All other things being equal, these states will also have a younger, and potentially safer, vehicle fleet. Therefore, it is possible that vehicle fleet average age could be biasing the light truck share coefficient, though the bias should be downward.

Table 3.4 presents evidence showing that the level of technology in a state's vehicle fleet is not downwardly biasing the coefficient on light truck share. Column (4) reports a regression of traffic fatalities on light truck share and a full set of controls and state and time effects.[22] Column (5) presents the same regression, but also includes as a control the ratio of vehicle sales over the past five years divided by total registered vehicles. The light truck share coefficient is 7 percent higher in column (5), suggesting that there is no downward bias from excluding the level of technology in a state's vehicle fleet.

A final specification test takes advantage of the fact that most of the growth in light truck share came after 1989, with the invention of sport utility vehicles. There is no reason to believe that underlying state demographic trends were substantially more pronounced during the 1990s than during the 1980s. However, the rate of growth in light truck share was 4.5 times higher post-1990 than in the 1980s. From 1981 to 1989, light truck share grew by an average of 0.29 percentage points per year, but from 1990 to 2004 light truck share grew by an average of 1.30 percentage points per year. The popularization of sport utility vehicles in 1990, with the release of the Ford Explorer, could therefore be conceptualized as a natural experiment which inflated the magnitude of changes in the light truck fleet. If demographic trends are correlated with changes in light truck share, and these trends do not become sharply more pronounced in the 1990s relative to the 1980s, then we would expect the light truck share coefficient post-1989 to be substantially lower in magnitude than

[22]This regression is identical to column (3) in Table 3.4 except that it is run on a shorter sample.

the light truck share coefficient pre-1990.

Columns (6) and (7) of Table 3.4 present regressions that test whether the light truck share coefficient falls after 1989. In column (6), the sample is restricted to the period from 1981 to 1989. In the column (7), the sample is restricted to the period from 1990 to 2004. The 1981-1989 light truck share coefficient is estimated at 0.297 while the 1990-2004 light truck share coefficient is estimated at 0.449, suggesting that unobserved state trends are not driving the light truck share coefficient.[23]

The results from state-level panel data suggest that a 10 percentage point increase in light truck share raises traffic fatalities by approximately 4.1 percent, or 1,720 deaths per year. However, it is unclear how much of this additional risk is borne by the occupants of the new light trucks and how much is borne by the occupants of other vehicles and pedestrians.

3.4 The Internal and External Distribution of Fatalities

3.4.1 Aggregate-Level Estimates

Section (3.3) presents state-level panel data results that quantify the effect of light trucks on total traffic fatalities but do not reveal how many of these fatalities accrue to light truck occupants and how many accrue to other users of the roadway system. One way to estimate this breakdown is to compare the overall fatality rates of light trucks and cars. For example, column (3) in Table 3.2 indicates that between 1996 and 2004, a one percentage point increase in light truck share raised fatalities by approximately 0.41 percent. During the same period, the occupant fatality rate for recent light trucks was 10.3 percent higher than for recent cars (Insurance Institute for Highway Safety, 2005; Ward's Automotive Yearbook, 1994-2004). If light trucks posed no additional risk to other roadway users, a 1 percentage point increase in light truck share would therefore raise total traffic fatalities by approximately 0.10 percent. This increase in fatalities would accrue exclusively to the occupants of new light trucks, so this figure represents an estimate of the internal fatality risk of light trucks. The remainder of the 0.41 percent total increase in traffic fatalities - 0.31 percent - represents an estimate of the external fatality risk, or the risk to other roadway users. Therefore, aggregate statistics suggest that internal fatalities account for approximately 25 percent

[23] Although the standard error of the pre-1990 coefficient is large, a 95 percent confidence interval would have an upper bound of 1.01, which is much less than 4.5 times the post-1990 coefficient.

of the total fatality increase and external fatalities account for the remaining 75 percent.

However, using aggregate statistics to estimate the breakdown of internal and external fatalities may generate misleading estimates if the average light truck or car driver is not representative of marginal light truck or car driver driver. In particular, Bradsher (2002) notes that the sport utility vehicle buyers tend to come from the safest group of drivers on the road: middle-aged, affluent men and women. If the marginal consumer, who is indifferent between a car and a light truck, is a safer driver than the average car buyer, then her fatality rate in a car will be lower than the average fatality rate of car drivers. In that case, using aggregate fatality rates to calculate the internal versus external fatality breakdown will underestimate the number of internal fatalities and overestimate the number of external fatalities. To address this problem, I examine accident-level micro data.

3.4.2 Empirical Framework

The contribution of a vehicle to the total number of traffic fatalities is equal to the sum of the vehicle's collision rate times the fatality rate in the collisions that it is involved in. A vehicle of type k therefore increases expected internal fatalities by

$$E[IF_k] = p_{cks} \cdot p_{fks} + p_{ckl} \cdot p_{fkl} + p_{ckh} \cdot p_{fkh} \tag{3.2}$$

where IF_k are internal fatalities, p_{cks} is the vehicle's collision involvement rate in single-vehicle collisions, p_{fks} is the vehicle's fatality rate in the event of a single-vehicle collision, and $p_{.kl}$ and $p_{.kh}$ are the same quantities for multi-vehicle collisions involving light vehicles and two-vehicle collisions involving heavy vehicles respectively. Likewise, a vehicle of type k increases expected external fatalities by

$$E[EF_k] = p_{ckl} \cdot p_{fkl} + p_{ckm} \cdot p_{fkm} + p_{ckp} \cdot p_{fkp} \tag{3.3}$$

where EF_k are external fatalities, p_{ckl} is the vehicle's collision involvement rate in multi-vehicle collisions, p_{fkl} is the fatality rate in other vehicles in the event of a multi-vehicle collision, and $p_{.km}$ and $p_{.kp}$ are the same quantities for light vehicle collisions involving motorcycles and pedestrians respectively.

To estimate fatality rates in the event of an accident - the $p_{f..}$ terms in equations (3.2) and (3.3) - I apply a methodology implemented in White (2004) to the accident-level micro data. I examine five types of accidents separately: two-vehicle collisions involving at least one car, two-vehicle collisions involving at least one light truck, single-vehicle collisions, two-vehicle collisions involving a motorcycle, and collisions involving a pedestrian or pedalcyclist.[24] For each accident type, I estimate a logit regression that specifies the probability of fatalities as a function of vehicle, driver, and accident characteristics.

For two-vehicle collisions, I define "vehicle 1" as the vehicle being struck. If the accident involves two vehicle types, then vehicle 1 is by default the vehicle type of that dataset (e.g., for two-vehicle collisions involving at least one car, vehicle 1 is always a car). If the accident involves two vehicles of the same type (i.e., two cars or two light trucks), then one of the vehicles is randomly assigned to be vehicle 1. The remaining vehicle I designate as "vehicle 2." The dependent variable is the presence of a fatality in vehicle 1, and the variable of interest is a dummy variable that takes a value of one if vehicle 2 is a light truck. For single-vehicle collisions and collisions involving pedestrians, the dependent variable is the presence of a fatality in the single vehicle or for the pedestrian, and the variable of interest is a dummy variable indicating whether the vehicle involved is a light truck.

All regressions are weighted so that the sample is representative of the population of accidents. Controls include the number of occupants inside the struck vehicle, vehicle model year, a dummy variable for heavy trucks, and dummy variables for medium cities, large cities, safety belt use (or helmet use, for motorcycles), rain, snow, fog, darkness, weekdays, negligent driving, drivers under the age of 21, drivers over the age of 60, interstate highways, divided highways, male drivers, young male drivers, and years.

3.4.3 Accident-Level Results

The first set of rows in Table 3.5 presents results for two-vehicle collisions involving cars. The first column reports the logit regression coefficient on the dummy variable indicating whether vehicle 2 is a light truck. The second column reports the probability of a fatality conditional on the vehicle type of the striking vehicle. For light trucks this probability is equal to the weighted average of the predicted logit regression probabilities taken across all accidents in which vehicle 2 is a light

[24]These five types of accidents account for approximately 85 percent of the accidents in the data set.

truck. For cars this probability is equal to the weighted average of the predicted logit regression probabilities taken across the same set of observations, but the vehicle 2 light truck indicator is set to zero. The probability comparison therefore fixes the covariates at the values associated with light truck drivers.[25] The third column reports the sample size for each data set.

The results for two-vehicle collisions involving cars indicate that light trucks pose a significant hazard to cars in the event of a collision. The light truck coefficient is positive and highly significant ($z = 3.59$). The probability of a fatality in vehicle 1 increases by 62 percent if vehicle 2 is a light truck rather than a car, from approximately 0.0011 to 0.0018.

The second set of rows in Table 3.5 presents results for two-vehicle collisions involving light trucks. These results indicate that light trucks also pose a significant hazard to other light trucks in the event of a collision. The light truck coefficient is positive and significant ($z = 3.09$). The probability of a fatality in vehicle 1 increases by 97 percent if vehicle 2 is a light truck rather than a car, from approximately 0.0007 to 0.0014. In fact, the probability of fatalities in a collision involving two light trucks is estimated to be 24 percent higher than in a collision involving two cars.

The third set of rows in Table 3.5 reports results for single-vehicle collisions. The light truck coefficient is again positive and significant ($z = 2.62$). The probability of a fatality in a single-vehicle collision increases by 20 percent, from approximately 0.0074 to 0.0089. A substantial portion of this increase may be due to increased rollover risk in light trucks; a logit regression with a rollover event as the dependent variable returns a light truck coefficient of 0.72 ($z = 40.53$).

The fourth and fifth sets of rows in Table 3.5 report results for two-vehicle collisions involving motorcyclists and single-vehicle collisions involving pedestrians or pedalcyclists. Light trucks increase the risk of fatalities in motorcycle collisions by 92 percent, from approximately 0.0377 to 0.0725 ($z = 3.85$ in the logit regression). They increase the probability of fatalities in pedestrian or pedalcyclist collisions by 77 percent, from approximately 0.0245 to 0.0434 ($z = 6.33$ in the logit regression).

Overall, light trucks pose a significant hazard to other vehicles in the event of a collision. Depending on the type of accident, the risk of fatality to others (or to themselves, in single-vehicle

[25] Alternatively, I could fix the covariates at the values associated with car drivers. However, because the state-level results are primarily identified off of car drivers who have become light truck drivers, it seems preferable to examine what would happen if those light truck drivers switched back to cars.

collisions) increases by 20 to 97 percent. However, these estimates may not be interpretable as causal effects in the traditional sense (e.g., Rubin, 1974).

3.4.4 Interpretation

The fatality process can be modeled as operating through both participation in accidents and the intensity of accidents. The probability of participation is the collision rate, or the $p_{c..}$ terms in equations (3.2) and (3.3). The intensity of accidents is the probability of a fatality in the event of a collision, or the $p_{f..}$ terms in equations (3.2) and (3.3). The fatality rates presented in the subsection above are estimates of intensity and are therefore conditional upon the event of an accident. However, Angrist (2001) notes that when the probability of participation is a function of intensity, estimates of intensity conditional on participation do not have a causal interpretation.

In this case, the primary concern is that light trucks may positively affect collision rates. If so, the conditional-on-collision difference in fatality rates between cars and trucks need not equal the difference in fatality rates between cars and trucks in a series of controlled crash tests. This discrepancy occurs because the set of accidents in which trucks are participating is different than the set of accidents in which cars are participating. Measuring the causal effect of crash intensity on fatalities requires us to observe cars participating in all the accidents in which trucks participate. Nevertheless, because the goal is to measure the distribution of external and internal fatalities rather than the causal effect of intensity on fatalities, the accident-level estimates are still useful. Equations (3.2) and (3.3) must hold because they are identities, even if the $p_{c..}$ terms do not have a causal interpretation.[26]

To compute collision rates, I examine the five types of collisions involving light duty vehicles (i.e., cars and light trucks) in the dataset: single-light vehicle, two-light vehicles, light vehicle-heavy truck, light vehicle-motorcycle, and light vehicle-pedestrian. Table 3.6 reports the annual collision rate for each type of collision; this rate is calculated by dividing the total number of collisions of each type by the total number of registered light vehicles.[27] Two-light vehicle collisions are the most

[26]Since we are tabulating internal and external fatalities, it is of no concern whether the increased fatality risk from light trucks arises because they are more dangerous in the same collision or because their characteristics cause them to enter into more dangerous collisions. Still of concern, however, is the possibility that driver characteristics of light trucks and cars are different and that these differences might affect fatality rates in the event of a collision. This concern is mitigated by the fact that controlling for a wide range of covariates, as is done in Table 3.5, does not affect the results very strongly.

[27]The total number of two-light vehicle collisions has been multiplied by two since each collision involves two light vehicles. Rates are calculated using vehicle registration data over the sample period from 1996 to 2004.

common, occurring at an annual rate of 0.027 per vehicle, and light vehicle-motorcycle collisions are the least common, occurring at an annual rate of 0.00014 per vehicle.

I first calculate the internal fatality effects of replacing 2.1 million light trucks with cars (one percent of the vehicle fleet) assuming that the collision rate of these vehicles remains unchanged. This is equivalent to changing the $p_{f..}$ terms in equation (3.2) while leaving the $p_{c..}$ terms unchanged.[28] There are three types of collisions involving internal fatalities: single-light vehicle, two-light vehicles, and light vehicle-heavy truck. Consider the internal fatality risk from single-vehicle accidents. The change in fatal crashes from switching 2.1 million light trucks to cars is the single-vehicle collision rate times the difference in fatality risk between cars and light trucks in single-vehicle collisions (from Table 3.5) times 2.1 million $(0.00638 \cdot (0.0074 - 0.0089) \cdot 2.1 \cdot 10^6)$. That number is multiplied by 1.15, as each fatal collision results in approximately 1.15 fatalities (White, 2004). The result, reported at the top of the first column in Table 3.7, implies that the switch would decrease single-vehicle collision fatalities by 23 per year if the collision rate remains unchanged. The two-light vehicle and light vehicle-heavy truck internal fatality changes are computed in a similar manner.[29] The fifth row in Table 3.7 totals the internal effects. The replacement of 2.1 million light trucks with cars is expected to increase internal fatalities by 3 deaths per year if the collision rate remains unchanged.

I then calculate the external fatality effects of replacing 2.1 million light trucks with cars (one percent of the vehicle fleet) assuming that the collision rate of these vehicles remains unchanged. This is equivalent to changing the $p_{f..}$ terms in equation (3.3) while leaving the $p_{c..}$ terms unchanged. There are three types of collisions that result in external fatalities: two-light vehicles, light vehicle-motorcycle, and light vehicle-pedestrian. Consider the external fatality risk from light vehicle-pedestrian accidents. The change in fatal crashes from switching 2.1 million light trucks to cars is the light vehicle-pedestrian collision rate times the difference in fatality risk between cars and light trucks in light vehicle-pedestrian collisions (from Table 3.5) times 2.1 million $(0.00046 \cdot (0.0245 - 0.0434) \cdot 2.1 \cdot 10^6)$. The result, reported in the third row of the first column in Table 3.7, implies that

[28] Note that in this scenario the difference in $p_{c..}$ terms would have a causal interpretation, because light trucks do not affect crash rates.

[29] For the two-light vehicle collisions, the fatality risk in a car conditional on a collision is computed as a weighted average of the fatality risk when a car is hit by a light truck and the risk when a car is hit by a car. The weights are 0.37 and 0.63 respectively, reflecting the average light truck share during the sample period. The analogous number for light trucks is calculated in a similar manner. For collisions involving heavy trucks, the fatality risk is computed using a logit regression run on a sample of two-vehicle collisions that involve one light vehicle and one heavy truck. The probability of a fatality is estimated at 0.0082 for a car and 0.0083 for a light truck.

the switch would decrease pedestrian fatalities by 18 per year if the collision rate remains unchanged. The two-light vehicle and light vehicle-motorcycle external fatality changes are computed in a similar manner.[30] The second-to-last row in Table 3.7 totals the external effects. The replacement of 2.1 million light trucks with cars is expected to decrease external fatal crashes by 75 per year if the collision rate remains unchanged.

The sum of the internal and external effects, reported in the last row of Table 3.7, is $3-75 = -72$, indicating that reducing the light truck share by one percentage point will reduce annual traffic fatalities by 72. However, estimates from the preferred state-level specification over the same sample period imply that a one percent reduction in light truck share should reduce annual traffic fatalities by 0.41 percent, or approximately 172. The discrepancy between these two estimates - 72 and 172 - implies that the collision rate of drivers who switch from light trucks to cars does not remain constant.

What collision rate differential between cars and light trucks can explain the discrepancy between the accident-level results and the state-level results? The second column in Table 3.7 reports the internal and external effects of replacing 2.1 million light trucks with cars under the assumption that the light truck crash rate is 30 percent higher than the car crash rate.[31] This is equivalent to assuming that each of the $p_{c..}$ terms in equations (3.2) and (3.3) is 30 percent higher for light trucks than for cars. This figure is implied by the data if we assume that the collision rate differential between light trucks and cars is constant across collision types; it is the unique parameter value that equalizes the state-level and accident-level estimates of total traffic fatalities.[32]

The estimates in the second column of Table 3.7 suggest that reducing the light truck share of the vehicle fleet by one percentage point reduces internal fatalities by 59 deaths per year and reduces

[30]For the two-light vehicle collisions, the fatality risk for a car conditional on a collision is computed as a weighted average between the fatality risk when a car is hit by a car and the risk when a light truck is hit by a car. The weights are 0.63 and 0.37 respectively, reflecting the average light truck share during the sample period. The final number is multiplied by 1.15.

[31]The estimates are computed in a similar manner to the first column, but the collision rates are specified such that the light truck collision rate is 30 percent higher than the car rate, and a weighted average of the two equals the overall collision rate. Also, the internal and external fatality risks in collisions involving two-light vehicles are updated to reflect the fact that the chance of being struck by a light truck is now higher.

[32]This assumption is likely to hold for collisions that only involve vehicles - for example, Levitt and Porter (2001) find that most driver characteristics, including drunk driving, increase collision risk in both single-vehicle and two-vehicle accidents by roughly equal proportions. However, the assumption may not hold for collisions involving pedestrians. I therefore examine how the estimates change if we allow the pedestrian collision rate differential between light trucks and cars to assume different values.

external fatalities by 114 deaths per year.[33] Therefore, external fatalities account for approximately 66 percent of the total change in traffic fatalities, and internal fatalities account for the remaining 34 percent. These estimates are reasonably similar to the breakdown based on aggregate fatality rates presented at the beginning of the section. Furthermore, the ratio of internal to external deaths is not very sensitive to variations in the relative pedestrian collision rate between cars and light trucks. For example, if cars and light trucks have the same pedestrian collision rate (rather than trucks having a collision rate that is 30 percent higher), then external fatalities account for 62 percent of the change in traffic fatalities. If light trucks have a pedestrian collision rate that is 50 percent higher than cars, then external fatalities account for 68 percent of the change in traffic fatalities.

The results in Table 3.7 and the estimates at the beginning of the section demonstrate that light trucks are a hazard not only to other drivers and pedestrians, but to their own occupants as well. The increase in internal fatality risk is primarily due to the higher estimated collision rate for light trucks. A variety of factors may contribute to this increased collision rate. First, light trucks generally have poor handling and braking in comparison to cars. Furthermore, their height relative to the roadway causes drivers to perceive themselves to be traveling slower than they are (Rist, 2001). Finally, if light truck drivers regard their vehicles as being more crashworthy, they may compensate by driving more aggressively (Peltzman, 1975). However, the first column in Table 3.7 shows that even if light trucks had the same collision rate as cars, their occupants would enjoy little additional protection on average, perhaps because of the increased rollover risk.

3.5 Conclusion

This paper analyzes the effects of light trucks on traffic fatalities by combining state-level and accident-level data sets. The results suggest that a one percentage point increase in the light truck share of the vehicle fleet will increase annual traffic fatalities by approximately 0.41 percent, or 172 deaths per year. Of these deaths, approximately one-quarter to one-third accrue to the occupants of the new light trucks, while the remaining two-thirds to three-quarters accrue to the occupants of other vehicles, pedestrians, and pedalcyclists. Using standard value of life estimates of seven million dollars per life, a vehicle life-span of fifteen years, and a real discount rate of three percent,

[33]These results suggest that the true effects are near the upper tail of the range presented in White (2004).

the implied Pigovian tax is approximately 4,650 dollars per light truck sold. Light trucks therefore pose a significant cost to other users of the highway system, but do not appear to provide any additional protection to their own occupants. Interpreted on a larger scale, the results suggest that the 21 percentage point increase in light truck share from 1981 to 2004 currently results in as many as 3,500 additional traffic fatalities per year. This estimate may partially explain why, despite advances in safety technology, the annual number of traffic fatalities in the United States has increased by almost 10 percent from 1992 to 2004, rising from 39,250 to 42,636.

Table 3.1: Summary Statistics

Variable	Mean	Std. Dev.
State Panel Data Set		
Traffic fatalities	880.8	876.6
Fatalities per 100,000 persons	18.4	6.2
Light truck share	0.296	0.114
Sample period	1981-2004	
Number of states	47	
Observations	1,128	
Accident-Level Data Set		
Average number of vehicles	1.80	0.66
Percent involving light trucks	0.470	0.499
Fatal accidents	5,717	
Total accidents	454,014	
Sample period	1996-2004	

Notes: Panel means are across state-time observations; they do not equal national means across time.

Table 3.2: Effect of Light Trucks on Traffic Fatalities

Variable:	(1)	(2)	(3)	(4)	(5)	(6)
Light truck share	0.414	0.362	0.415	0.420	0.345	0.297
	(0.151)	(0.150)	(0.236)	(0.373)	(0.153)	(0.153)
Population	1.03	0.96	1.00	0.66	0.98	0.94
	(0.10)	(0.09)	(0.28)	(0.18)	(0.11)	(0.09)
Controls	Full Set	Base Set	Full Set	None	Full Set	Base Set
Estimation method	Neg Bin	Neg Bin	Neg Bin	Neg Bin	WLS	WLS
Sample period	1981-2004	1981-2004	1996-2004	1981-2004	1981-2004	1981-2004
Observations	1,128	1,128	423	1,128	1,128	1,128

Notes: Parentheses contain standard errors clustered by state. All regressions include state and year effects and are weighted by state population.

Table 3.3: Vehicle Fleet Endogeneity

	(1)	(2)	(3)	(4)
Dependent variable	Log Truck Sales Share	Truck Sales Share	Truck Sales Share	Log Fatalities
Right-hand variable	Truck Share	Log Truck Share	Truck Share	Truck Share
Light truck share				0.375
				(0.162)
Fatalities per capita	0.011	0.010	0.098	
	(0.012)	(0.037)	(0.073)	
Fatalities per capita (t-1)	0.011	0.024	0.079	
	(0.010)	(0.030)	(0.064)	
Fatalities per capita (t-2)	0.005	0.003	0.047	
	(0.006)	(0.020)	(0.038)	
Fatalities per capita (t-3)	-0.001	-0.028	0.036	
	(0.007)	(0.022)	(0.045)	
Fatalities per capita (t-4)	-0.007	-0.008	0.013	
	(0.007)	(0.022)	(0.036)	
Sum of coefficients	0.019	0.001	0.273	
Sample period	1985-2001	1985-2001	1985-2001	1985-2001
Observations	799	799	799	799

Notes: Parentheses contain standard errors clustered by state. All regressions include a full set of controls, state and year effects, and are weighted by state population.

Table 3.4: Measurement Error and Unobserved State Trends

Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Light truck share	0.414	0.401	0.496	0.515	0.550	0.297	0.449
	(0.151)	(0.164)	(0.160)	(0.152)	(0.171)	(0.367)	(0.171)
Population	1.03	0.65	1.33	1.62	1.34	0.52	1.44
	(0.10)	(0.14)	(0.29)	(0.31)	(0.22)	(0.16)	(0.15)
VMT		0.46					
		(0.10)					
New vehicle share					0.253		
					(0.165)		
State time trends	No	No	Yes	Yes	Yes	No	No
Sample period	1981-2004	1981-2004	1981-2004	1981-2001	1981-2001	1981-1989	1990-2004
Observations	1,128	1,128	1,128	987	987	423	705

Notes: Dependent variable in all regressions is traffic fatalities. Parentheses contain standard errors clustered by state. All regressions include a full set of controls, state and year effects, and are weighted by state population.

Table 3.5: Predicted Probabilities of Fatality in Struck Vehicle

	Logit Results	Fatality Probability	Sample
Two-vehicle crashes w/cars			190,791
Veh. 2 is light truck	0.486	0.00178	
	(0.135)		
Veh. 2 is car		0.00110	
Two-vehicle crashes w/light trucks			117,549
Veh. 2 is light truck	0.684	0.00137	
	(0.221)		
Veh. 2 is car		0.00070	
Single-vehicle crashes			118,069
Veh. is light truck	0.191	0.00891	
	(0.073)		
Veh. is car		0.00741	
Crashes w/motorcycles			3,191
Veh. 2 is light truck	0.728	0.07245	
	(0.189)		
Veh. 2 is car		0.03773	
Crashes w/pedestrians or pedalcyclists			15,990
Veh. is light truck	0.663	0.04342	
	(0.105)		
Veh. is car		0.02451	

Notes: Parentheses contain standard errors. All regressions include controls and are weighted by sampling weights.

Table 3.6: Annual Per Light Vehicle Collision Rates

Collision Type:	Annual Rate
Two light vehicles	0.02747
Single light vehicle	0.00638
Light vehicle and heavy truck	0.00127
Light vehicle and motorcycle	0.00014
Light vehicle and pedestrian or pedalcyclist	0.00046

Notes: Rates are over 1996-2004 period.

Table 3.7: Effects of Replacing 2.1 Million Light Trucks with Cars

	(1)	(2)
Ratio of Light Truck Collision Rate to Car Collision Rate	1.00	1.30
Internal Effects:		
Single-vehicle collisions	-23	-58
Two-vehicle collisions	27	6
Light vehicle-heavy truck collisions	0	-7
Total Internal Deaths	3	-59
External Effects:		
Two-vehicle collisions	-46	-71
Collisions with motorcycles	-10	-15
Collisions with pedestrians/pedalcyclists	-18	-28
Total External Deaths	-75	-114
Net Effect:	-72	-173

Notes: Numbers may not sum exactly due to rounding.

Bibliography

Abadie, Alberto, Joshua Angrist, and Guido Imbens (2002) 'Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings.' *Econometrica* 70(1), 91–117

Adda, Jerome, Tarani Chandola, and Michael Marmot (2003) 'Socio-economic Status and Health: Causality and Pathways.' *Journal of Econometrics* 112(1), 57–63

Allison, Paul, and Richard Waterman (2002) 'Fixed Effects Negative Binomial Regression Models.' In *Sociological Methodology 2002,* ed. Ross Stolzenberg (Basil Blackwell)

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)* (American Psychiatric Association)

Angrist, Joshua (2001) 'Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors: Simple Strategies for Empirical Practice.' *Journal of Business and Economic Statistics* 19(1), 2–16

Angrist, Joshua, and Alan Krueger (1991) 'Does Compulsory School Attendance Affect Schooling and Earnings?' *Quarterly Journal of Economics* 106(4), 979–1014

—— (1995) 'Split-Sample Instrumental Variables Estimates of the Return to Schooling.' *Journal of Business and Economic Statistics* 13(2), 225–35

Angrist, Joshua, and Guido Imbens (1995) 'Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity.' *Journal of the American Statistical Association* 90(430), 431–42

Angrist, Joshua, and Kevin Lang (2004) 'Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program.' *American Economic Review* 94(5), 1613–1634

Angrist, Joshua, Guido Imbens, and Alan Krueger (1999) 'Jackknife Instrumental Variables Estimation.' *Journal of Applied Econometrics* 14(1), 57–67

Angrist, Joshua, Guido Imbens, and Donald Rubin (1996) 'Identification of Causal Effects Using Instrumental Variables.' *Journal of the American Statistical Association* 91(434), 444–55

Ashenfelter, Orley, and Michael Greenstone (2004) 'Using Mandated Speed Limits to Measure the Value of a Statistical Life.' *Journal of Political Economy* 112(1), S226–S267

Australian Institute of Family Studies (2005) 'Growing Up in Australia: The Longitudinal Study of Australian Children: 2004 Annual Report.' Australian Institute of Family Studies

Autor, David, and Susan Houseman (2005) 'Do Temporary Help Jobs Improve Labor Market Outcomes for Low-Skilled Workers? Evidence from Random Assignments.' MIT Working Paper

Berreuta-Clement, J., L. Schweinhart, W. S. Barnett, A. Epstein, and D. Weikart (1984) *Changed Lives: The Effects of the Perry Preschool Program on Youths Through Age 19* (High/Scope Press)

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004) 'How Much Should We Trust Difference in Difference Estimates?' *Quarterly Journal of Economics* 119(1), 249–275

Bosma, H., M. G. Marmot, H. Hemingway, A. C. Nicholson, E. Brunner, and S. A. Stansfeld (1997) 'Low Job Control and Risk of Coronary Heart Disease in Whitehall II (Prospective Cohort) Study.' *British Medical Journal* 314(7080), 558–65

Bound, John, David Jaeger, and Regina Baker (1995) 'Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak.' *Journal of the American Statistical Association* 90(430), 443–450

Bradsher, Keith (2002) *High and Mighty: SUVs: The World's Most Dangerous Vehicles and How They Got That Way* (PublicAffairs)

Campbell, Frances, and Craig Ramey (1994) 'Effects of Early Intervention on Intellectual and Academic Achievement: A Follow-Up Study of Children from Low-Income Families.' *Child Development* 65(2), 684–698

Campbell, Frances, Craig Ramey, Elizabeth Pungello, Joseph Sparling, and Shari Miller-Johnson (2002) 'Early Childhood Education: Young Adult Outcomes From the Abecedarian Project.' *Applied Developmental Science* 6(1), 42–57

Campbell, Frances, Elizabeth Pungello, Shari Miller-Johnson, Margaret Burchinal, and Craig Ramey (2001) 'The Development of Cognitive and Academic Abilities: Growth Curves From an Early Childhood Educational Experiment.' *Developmental Psychology* 37(2), 231–242

Carneiro, Pedro, and James Heckman (2003) 'Human Capital Policy.' In *Inequality in America: What Role for Human Capital Policies?,* ed. James Heckman and Alan Krueger (MIT Press)

Chandola, T., J. Siegrist, and M. G. Marmot (2005) 'Do Changes in Effort-Reward Imbalance at Work Contribute to an Explanation of the Social Gradient in Angina?' *Occupational and Environmental Medicine* 62(4), 223–30

Chandola, Tarani, Mel Bartley, Amanda Sacker, Crispin Jenkinson, and Michael Marmot (2003) 'Health Selection in the Whitehall II Study, UK.' *Social Science and Medicine* 56(10), 2059–72

Clarke, Stevens, and Frances Campbell (1998) 'Can Intervention Early Prevent Crime Later? The Abecedarian Project Compared with Other Programs.' *Early Childhood Research Quarterly* 13(2), 319–343

Coate, Douglas, and James VanderHoff (2001) 'The Truth About Light Trucks.' *Regulation* 24(1), 22–27

Crandall, Robert, and John Graham (1989) 'The Effect of Fuel Economy Standards on Automobile Safety.' *Journal of Law and Economics* 32(1), 97–118

Cunha, Flavio, James Heckman, Lance Lochner, and Dimitriy Masterov (2005) 'Interpreting the Evidence on Life Cycle Skill Formation.' NBER Working Paper Series, Working Paper 11331

Currie, Janet (2001) 'Early Childhood Education Programs.' *Journal of Economic Perspectives* 15(2), 213–238

Currie, Janet, and Duncan Thomas (1995) 'Does Head Start Make a Difference?' *American Economic Review* 85(3), 341–364

—— (2000) 'School Quality and the Longer-Term Effects of Head Start.' *Journal of Human Resources* 35(4), 755–774

Currie, Janet, and Matthew Neidell (2004) 'Getting Inside the "Black Box" of Head Start Quality: What Matters and What Doesn't.' UCLA Department of Economics, manuscript

Deaton, Angus (2003) 'Health, Inequality, and Economic Development.' *Journal of Economic Literature* 41(1), 113–158

Deaton, Angus, and Christina Paxson (1999) 'Mortality, Education, Income and Inequality Among American Cohorts.' NBER Working Paper Series, Working Paper 7140

Dee, Thomas (2005) 'A Teacher Like Me: Does Race, Ethnicity or Gender Matter?' *American Economic Review Papers and Proceedings* 95(2), 158–165

Donald, Stephen, and Kevin Lang (2004) 'Inference with Difference in Differences and Other Panel Data.' Boston University Department of Economics, manuscript

Evans, Leonard (1985) 'Involvement Rate in Two-Car Crashes Versus Driver Age and Car Mass of Each Involved Car.' *Accident Analysis and Prevention* 17(2), 155–70

—— (2001) 'Causal Influence of Car Mass and Size on Driver Fatality Risk.' *American Journal of Public Health* 91(7), 1076–81

Evans, Leonard, and Michael Frick (1993) 'Mass Ratio and Relative Driver Fatality Risk in Two-Vehicle Crashes.' *Accident Analysis and Prevention* 25(2), 213–24

—— (1994) 'Car Mass and Fatality Risk: Has the Relationship Changed?' *American Journal of Public Health* 84(1), 33–36

Evans, William, and John Graham (1988) 'Traffic Safety and the Business Cycle.' *Alcohol, Drugs, and Driving* 4(1), 31–38

Figlio, David (2005) 'Boys Named Sue: Disruptive Children and their Peers.' National Bureau of Economic Research Working Paper No. 11277

Fredriksson, Peter, and Björn Öckert (2005) 'Is Early Learning Really More Productive? The Effect of School Starting Age on School and Labor Market Performance.' IZA Discussion Paper Series, No. 1659

Garces, Eliana, Duncan Thomas, and Janet Currie (2002) 'Longer-Term Effects of Head Start.' *American Economic Review* 92(4), 999–1012

Gayer, Ted (2004) 'The Fatality Risks of Sport-Utility Vehicles, Vans, and Pickups Relative to Cars.' *The Journal of Risk and Uncertainty* 28(2), 103–33

Gray, Susan, and Rupert Klaus (1965) 'An Experimental Preschool Program for Culturally Deprived Children.' *Child Development* 36(4), 887–898

—— (1970) 'The Early Training Project: A Seventh-Year Report.' *Child Development* 41(4), 909–924

Gray, Susan, Barbara Ramsey, and Rupert Klaus (1982) *From 3 to 20: The Early Training Project* (University Park Press)

Hanushek, Eric (1996) 'School Resources and Student Performance.' In *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success.*, ed. Gary Burtless (Brookings Institution)

—— (2003) 'Comments.' In *Inequality in America: What Role for Human Capital Policies?,* ed. James Heckman and Alan Krueger (MIT Press)

Hausman, Jerry, Bronwyn Hall, and Zvi Griliches (1984) 'Econometrics Models for Count Data with an Application to the Patents-R & D Relationship.' *Econometrica* 52(4), 909–38

Heckman, James (2005) 'Invited Comments.' In *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40,* ed. L. Schweinhart, et al. (High/Scope Press)

Heckman, James, and Yonah Rubinstein (2001) 'The Importance of Noncognitive Skills: Lessons from the GED Testing Program.' *American Economic Review* 91(2), 145–149

Her Majesty's Stationery Office (1971) 'Civil service statistics 1971.' Her Majesty's Stationery Office

HM Treasury Office (1983) 'Civil service statistics 1983.' Her Majesty's Stationery Office

—— (1985) 'Civil service statistics 1985.' Her Majesty's Stationery Office

—— (1986) 'Civil service statistics 1986.' Her Majesty's Stationery Office

Horowitz, Joel (2001) 'The Bootstrap in Econometrics.' In *Handbook of Econometrics,* ed. James Heckman and Edward Leamer, vol. 5 (Elsevier Science B.V.) chapter 52, pp. 3159–3228

Hoyert, Donna, Kenneth Kochanek, and Sherry Murphy (1999) 'Deaths: Final Data for 1997.' *National Vital Statistics Reports*

Inoue, Atsushi, and Gary Solon (2005) 'Two-Sample Instrumental Variables Estimators.' National Bureau of Economic Research Technical Working Paper No. 311

Insurance Institute for Highway Safety (2005) 'Fatality Facts 2004: Occupants of Cars, Pickups, SUVs, and Vans.' http://www.iihs.org/

Joksch, Hans (1998) 'Fatality risks in collisions between cars and light trucks.' National Highway Traffic Safety Administration

Keeler, Theodore (1994) 'Highway Safety, Economic Behavior, and Driving Environment.' *American Economic Review* 84(3), 684–93

Kirp, David (2005) 'All My Children.' *The New York Times,* July 31, 2005, Section 4A, 20

Kling, Jeffrey, and Jeffrey Liebman (2004) 'Experimental Analysis of Neighborhood Effects on Youth.' Kennedy School of Government Working Paper No. RWP04-034

Krueger, Alan (1999) 'Experimental Estimates of Education Production Functions.' *The Quarterly Journal of Economics* 114(2), 497–532

—— (2003) 'Inequality, Too Much of a Good Thing.' In *Inequality in America: What Role for Human Capital Policies?,* ed. James Heckman and Alan Krueger (MIT Press)

Kuper, H., and M. G. Marmot (2003) 'Job Strain, Job Demands, Decision Latitude, and Risk of Coronary Heart Disease Within the Whitehall II Study.' *Journal of Epidemiology and Community Health* 57(2), 147–53

Levitt, Steve, and Jack Porter (2001) 'How Dangerous Are Drinking Drivers?' *Journal of Political Economy* 109(6), 1198–1237

Lleras-Muney, Adriana (2005) 'The Relationship Between Education and Adult Mortality in the United States.' *Review of Economic Studies* 72(1), 189–221

Marmot, M. G., G. D. Smith, S. Stansfeld, C. Patel, F. North, J. Head, I. White, E. Brunner, and A. Feeney (1991) 'Health Inequalities Among British Civil Servants: The Whitehall II Study.' *The Lancet* 337(8754), 1387–93

Marmot, M. G., Geoffrey Rose, M. Shipley, and P. J. S. Hamilton (1978) 'Employment Grade and Coronary Heart Disease in British Civil Servants.' *Journal of Epidemiology and Community Health* 32(4), 244–9

Marmot, M. G., H. Bosma, H. Hemingway, E. Brunner, and S. Stansfeld (1997) 'Contribution of Job Control and Other Risk Factors to Social Variations in Coronary Heart Disease Incidence.' *The Lancet* 350(9073), 235–39

Marmot, M. G., M. J. Shipley, and Geoffrey Rose (1984) 'Inequalities in Death - Specific Explanations of a General Pattern?' *The Lancet* 1(8384), 1003–6

Marmot, Michael (2003) 'Understanding Social Inequalities in Health.' *Perspectives in Biology and Medicine* 46(3), S9–S23

Marmot, Michael, and Eric Brunner (2004) 'Cohort Profile: The Whitehall II Study.' *International Journal of Epidemiology*

Marmot, Michael, and George Davey Smith (1997) 'Socio-economic Differentials in Health: The Contribution of the Whitehall Studies.' *Journal of Health Psychology* 2(3), 283–96

Mazumder, Bhashkar (2005) 'Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data.' *Review of Economics and Statistics* 87(2), 235–55

Mealli, Fabrizia, and Donald Rubin (2003) 'Assumptions Allowing the Estimation of Direct Causal Effects.' *Journal of Econometrics* 112(1), 79–87

Michener, Ron, and Carla Tighe (1992) 'A Poisson Regression Model of Highway Fatalities.' *American Economic Review Papers and Proceedings* 82(2), 452–56

Miller, Jerome (1992) 'Hobbling a Generation: Young African American Males in the Criminal Justice System of America's Cities: Baltimore, Maryland.' National Center on Institutions and Alternatives

Miskie, Brooke, Susan Near, and Robert Hegele (2003) 'Comment on Survival in Academy Award-Winning Actors and Actresses.' *The Annals of Internal Medicine* 138(1), 77–8

National Highway Traffic Safety Administration (2006) 'Five Star Crash Test and Rollover Ratings.' http://safercar.gov/

O'Brien, Peter (1984) 'Procedures for Comparing Samples with Multiple Endpoints.' *Biometrics* 40(4), 1079–1087

Oden, S., L. Schweinhart, D. Weikart, S. Marcus, and Y. Xie (2000) *Into Adulthood: A Study of the Effects of Head Start* (High/Scope)

Peltzman, Sam (1975) 'The Effect of Automobile Safety Regulation.' *Journal of Political Economy* 83(4), 677–725

Ramey, Craig, Keith Yeates, and Elizabeth Short (1984) 'The Plasticity of Intellectual Development: Insights from Preventative Intervention.' *Child Development* 55(5), 1913–1925

Redelmeier, Donald, and Sheldon Singh (2001) 'Survival in Academy Award-Winning Actors and Actresses.' *Annals of Internal Medicine* 134(10), 955–62

Reid, D. D., G. Z. Brett, P. J. S. Hamilton, R. J. Jarrett, Harry Keen, and Geoffrey Rose (1974) 'Cardiorespiratory Disease and Diabetes Among Middle-Aged Male Civil Servants.' *The Lancet* 1(7856), 469–73

Rist, Curtis (2001) 'Roll Over, Newton: The Design of Sport Utility Vehicles Is Enough to Make the Father of Physics Turn in His Grave.' *Discover* 22(4), 44–49

Rubin, Donald (1974) 'Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.' *Journal of Educational Psychology* 66(5), 688–701

Ruhm, Christopher (1995) 'Economic Conditions and Alcohol Problems.' *Journal of Health Economics* 14(5), 583–603

___ (2000) 'Are Recessions Good For Your Health?' *Quarterly Journal of Economics* 115(2), 617–650

Schweinhart, L., H. Barnes, and D. Weikart (1993) *Significant Benefits: The High/Scope Perry Preschool Study Through Age 27* (High/Scope Press)

Schweinhart, L., J. Montie, Z. Xiang, W. S. Barnett, C. Belfield, and M. Nores (2005) *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40* (High/Scope Press)

Shively, Carol, and Thomas Clarkson (1994) 'Social Status and Coronary Artery Atherosclerosis in Female Monkeys.' *Atherosclerosis and Thrombosis* 14(5), 721–26

Simon, Julian (1997) *Resampling: The New Statistics* (Resampling Stats)

Smith, James (1999) 'Healthy Bodies and Thick Wallets: The Dual Relation Between Health and Economic Status.' *Journal of Economic Perspectives* 13(2), 145–166

Solon, Gary (1992) 'Intergenerational Income Mobility in the United States.' *American Economic Review* 82(3), 393–408

Stanley, Martin (2004) *How To Be A Civil Servant* (Politico's Publishing)

United States Census Bureau (2005) 'International Data Base Summary Demographic Data'

U.S. Department of Transportation, Federal Highway Administration (1981-2004) 'Highway statistics series.' U.S. Government Printing Office

Viscusi, W. Kip, and Joseph Aldy (2003) 'The Value of a Statistical Life: A Critical Review of Market Estimates throughout the World.' *Journal of Risk and Uncertainty* 27(1), 5–76

Ward's Reports Incorporated (1980-2005) *Ward's Automotive Yearbook* (Ward's Reports Incorporated)

Weikart, D., D. Deloria, S. Lawser, and R. Wiegerink (1970) *Longitudinal Results of the Ypsilanti Perry Preschool Project* (High/Scope Press)

Westfall, Peter, and S. Young (1993) *Resampling-Based Multiple Testing* (John Wiley and Sons)

White, Michelle (2004) 'The "Arms Race" on American Roads: The Effect of Sport Utility Vehicles and Pickup Trucks on Traffic Safety.' *Journal of Law and Economics* 47(2), 333–55

Wilkinson, Richard, and Kate Pickett (2005) 'Income Inequality and Population Health: A Review and Explanation of the Evidence.' *Social Science and Medicine.* (available online 13 October 2005)

Yucesan, Enver (1995) 'Using Nonparametric Statistics in Simulation Analysis: A Review.' In *Proceedings of the 1995 Winter Simulation Conference,* ed. C. Alexopoulos, K. Kang, W. Lilegdon, and D. Goldsman pp. 141–146

Zimmerman, David (1992) 'Regression Toward Mediocrity in Economic Stature.' *American Economic Review* 82(3), 409–29