# INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

**The quality of this reproduction is dependent upon the quality of the copy submitted.** Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

# Empirical Essays on Incentives, Firm Coordination, and Social Spillovers

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Stephen Emmanuel Cacciola

Dissertation Director: Michael A. Boozer

December 2002

UMI Number: 3068256





ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Empirical Essays on Incentives, Firm Coordination, and Social Spillovers

Stephen Emmanuel Cacciola

2002

This dissertation consists of three chapters that study how the allocation of resources within organizations affects individual and firm performance. Chapter 1 analyzes a recent monitoring technology in the trucking industry, on-board computers (OBCs), which has the capacity to both improve employee incentives and enhance resource allocation (firm coordination) decisions. The empirical methodology employed improves on previous evaluations of this technology by looking at the direct impact of adoption on measures that incorporate incentive and coordination effects. The results indicate that the incentive effect of OBCs is manifested in firms' operating costs, particularly maintenance costs, as truckers who are monitored drive their trucks in a manner that is preferred by their firms. The coordination benefits appear to enable adopting firms to substantially increase the size (revenue) of their operations. Chapter 2 exploits the incentive capabilities of OBCs to undertake rigorous empirical testing of a multitask principal-agent model. Tests of agency theory are provided along two dimensions: first, do observed contracts vary as predicted by the theory?, and second, to what extent do employees respond to changes in incentives? Tentative evidence is found that more precise measurement of driver behavior is associated with a greater incidence of performance bonuses. More compelling evidence is provided that a better means of monitoring leads to a substantial improvement in employee effort, as measured by increases in truck life expectancy and fuel efficiency. Chapter 3, co-authored with Michael Boozer, examines the allocation of students to classes in elementary schools. We propose an estimation scheme for endogenous peer group effects (i.e. spillover or feedback effects) that utilizes a randomly assigned social program that operates at differing intensities within and between peer groups. The data used are from Project STAR, a class size reduction experiment conducted in Tennessee. We argue that the

Small class treatment itself created class groupings of varying quality. When allowance is made for this feedback effect of prior exposure to the Small class treatment, it is found that the peer effects account for much of the total experimental effects in the later grades, and the direct class size effects are rendered substantially smaller.

**Acknowledgments**

I am indebted to several individuals for their assistance and contributions to this work. Many thanks to Michael Boozer for countless valuable discussions and his hand in guiding my progress. Paul Schultz and Ann Stevens provided extremely helpful advice and comments at all stages of this project. I also wish to thank Panle Jia and Tavneet Suri for comments on earlier drafts, as well as participants at seminars, particularly those at the Yale Labor and Population Workshop, for their remarks. Finally, special thanks to my family for their constant support and encouragement.

# Contents

# List of Figures

# List of Tables

Chapter 1

## Introduction to the Dissertation

This dissertation consists of three chapters that study, broadly speaking, how the allocation of resources within organizations affects individual and firm performance. Chapters 1 and 2 consider a recent technological innovation in the trucking industry. These chapters analyze how the implementation of a monitoring device can improve worker efficiency and firm decision-making, as well as alter the nature of the relationship between employers and employees. Chapter 3 examines the allocation of students to classes in elementary schools. In particular, the existence of endogenous peer group effects (and, therefore, social spillovers) implies that this allocation decision can have a substantial impact on student outcomes.

Modern theories of the firm emphasize that the structure and performance of organizations is in part determined by their ability to implement sophisticated new technologies. Chapter 1 studies a recent monitoring technology in the trucking industry, on-board computers (OBCs), which has the capacity to both improve employee incentives and enhance resource allocation (and firm coordination) decisions. The empirical methodology used here improves on the previous evaluations of this technology by looking at the *direct* impact of adoption on measures that incorporate incentive and coordination effects. The results indicate that the incentive effect of OBCs is manifested in firms' operating costs, particularly outside maintenance costs, as truckers who are monitored drive their trucks in a manner that is preferred by their firms. The coordination benefits, broadly defined to include the superior customer service afforded by the technology, enable adopting firms to increase the size of their operations. The estimates imply that carriers that have outfitted their entire fleets of trucks with OBCs which possess both the incentive and coordination features have boosted their revenues on the order of 40% to 50%.

Chapter 2 exploits the incentive capabilities of OBCs to test one of the workhorse models

of microeconomics. While the theoretical underpinnings of personnel economics and agency theory have been richly developed, rigorous empirical testing and evaluation of these models has lagged far behind. Chapter 2 contributes to the empirical evidence in this field by providing tests of agency theory along two dimensions: first, do observed contracts vary as predicted by the theory?, and second, to what extent do employees respond to changes in incentives? Tentative evidence is found that more precise measurement of truck driver behavior is associated with a greater incidence of performance bonuses, a central prediction of the theory. More compelling evidence is provided that a better means of monitoring leads to a substantial improvement in employee effort. In the trucking context, changes in driver behavior stemming from the adoption of an OBC lead to roughly a one year increase in truck life expectancy and a 3% upgrade in fuel efficiency.

Chapter 3, co-authored with Michael Boozer, proposes an estimation scheme that identifies endogenous peer group effects, i.e. spillover or feedback effects. We argue that such effects are most credibly identified by a randomly assigned social program that operates at differing intensities within and between peer groups. The data used are from Project STAR, a class size reduction experiment conducted in Tennessee elementary schools. In these data, classes are comprised of varying fractions of students who had previously been exposed to the Small class treatment, creating class groupings of varying experimentally induced quality. This variation in class group quality is used to estimate the spillover effect. When allowance is made for this feedback effect of prior exposure to the Small class treatment, it is found that the peer effects account for much of the total experimental effects in the later grades, and the direct class size effects are rendered substantially smaller.

# Chapter 1
# The Impact of a Monitoring Technology on Worker Incentives and the Coordination of Firm Activity: Evidence from the Trucking Industry

## 1 Introduction

The development and implementation of sophisticated new technologies have tremendous implications for how economic activity is structured within organizations. Consider the enhanced ability of firms to monitor how workers perform their jobs. Starting with Alchian and Demsetz (1972), economists have recognized the potential of monitoring to improve employee incentives and prevent shirking. Holmstrom (1979, 1982) and Holmstrom and Milgrom (1994), among others, emphasize the role that the precision in measuring employee actions plays in structuring efficient contracts. The field of personnel economics has built on these ideas, identifying testable implications of theories for applied economists and providing more guidance to real-world practitioners involved in making business decisions.[1] Certain technologies also allow managers who provide strategic direction and their subordinates involved in production to communicate more easily with each other. While the economic literature regarding how information is used within organizations is more diffuse, it is clear that this type of technology influences how work is delegated and the extent to which production decisions are centralized.[2] In particular, a more fluid flow of information across economic actors, if used efficiently, should improve resource allocation decisions and how

[1] For surveys of issues in personnel economics, see Prendergast (1996, 1999) and Lazear (1995, 1999, 2000).

[2] The literature on the efficient use of information effectively starts with the classic work of Hayek (1945). More recently, a vast literature has developed that studies the contributions of computers and information technology to business. See Brynjolfsson and Hitt (2000) for a survey.

activities are coordinated within a firm.

This paper analyzes a monitoring technology in the trucking industry that incorporates both *incentive* and *coordination* enhancing capabilities. Starting in the late 1980's, on-board computer (OBC) technology became available, whereby small computers could be installed on individual trucks. Additionally, there are two classes of OBCs, each providing somewhat different monitoring features. A trip recorder keeps a running electronic log of how a truck is operated. When the driver returns back to the firm after making one or more hauls, the trip recorder's contents are downloaded to the firm's computers, where the data are processed and analyzed. As such, trip recorders are valuable for improving driver *incentives*. Electronic Vehicle Management Systems (EVMS), on the other hand, provide trucking firms with all of the information that can be obtained by trip recorders, as well as a real-time exchange of information, including truck location, between drivers and firms. This additional information can be used by firms to more efficiently assign trucks and drivers to hauls, improving resource allocation and the *coordination* of firm activity. Earlier work by Hubbard (2000) has studied the adoption of this new technology, but due to data limitations has had to rely on rather indirect inferences as to the effects of OBCs.[3] By using data sets with information on OBC adoption and firm-level financial variables, I provide *direct* estimates of the value to firms of the incentive and coordination capabilities, which I separately identify by looking at the impact of adoption on measures associated with these features.

The empirical work in this paper has two sections. In the first I reanalyze the work

[3]A number of other papers have used the introduction of OBCs in trucking to study a variety of issues, including the interplay of information technology and marketing objectives, asset ownership, the 'make versus buy' decision, the effect of technology on worker's lives, capacity utilization, and contract choice. See Chakraborty and Kazarosian (1999), Baker and Hubbard (2000), Baker and Hubbard (2001), Belman and Monaco (2001), Hubbard (2001), and Cacciola (2002).

of Hubbard (2000), which estimates the value of OBCs by comparing the use rates of trip recorders and EVMS across different sectors of the trucking industry. I improve on the empirical implementation of this strategy by more carefully considering the variation in OBC adoption. I argue that in order to separate the incentive and coordination effects, one must take account of the differential replacement rate of trucks by sector. New trucks tend to come bundled with the current form of OBC, leading to a secular trend of EVMS use across model years. To factor out this secular increase in EVMS adoption, I show that we need to rely on the *within* model year variation in OBC use, as the *between* model year variation is corrupted by truck age and replacement. After this adjustment is made, the incentive effect of OBCs remains, while the coordination effect becomes very difficult to detect.

After establishing the fragility of this approach, a following section implements a strategy that more directly estimates the incentive and coordination effects, providing easily interpretable measures of their benefits. In particular, I use the fact that the incentive effect manifests itself in how drivers operate trucks. Firms care about this type of driver behavior because poor driving technique reduces fuel efficiency and causes wear and tear on the truck. A technology that monitors the driver can help alter his behavior by allowing for a compensation contract that is contingent on his actions. Since incentive improvements fundamentally influence the truck itself, I look at the impact of OBC adoption on several measures of truck operating *costs*. In order to assess a causal effect, I implement a Deaton (1985) synthetic panel data approach by creating cell-level groupings that combine OBC data and firm-level financial data. The empirical work indicates that adoption does reduce costs, particularly the costs spent on the maintenance of trucks performed by mechanics outside the firm. The coordination effect, defined broadly to include the improved quality of customer service via shipment tracking capabilities, is largely scale-enhancing, supplying

firms the means to schedule additional shipments. To capture this feature, I look at the impact of EVMS adoption on firm *revenue*. The estimates, using both cross section and fixed effect specifications, indicate large effects of adoption consistent with the hypothesized coordination effect.

The remainder of the paper proceeds as follows. Section 2 provides some institutional detail of the trucking industry, describes how OBCs may affect incentives and coordination, and sketches out the comparisons a firm uses in deciding whether or not to adopt a technology. I also discuss disparate means of empirically estimating the value of incentive and coordination improvements from adoption. Section 3 reanalyzes the previous approach used in the literature, emphasizing the contribution of truck replacement to the variation in adoption. Section 4 presents a framework for more directly evaluating the technology, and presents the empirical results of the cost and revenue analyses. Finally, Section 5 concludes.

# 2 Organizational Issues in the Trucking Industry

## 2.1 Institutions

The trucking industry is characterized by considerable diversity in its operations and services. There exist fairly distinct, though not completely independent, sectors within the industry which serve as a means for firms to offer varied products to their customers.[4] Trucking firms, called 'carriers', own transportation equipment such as truck-tractors and trailers, and provide their services to 'shippers' who need goods moved from one place to another. Carriers are divided into two broad categories. 'For-hire' firms offer to move cargo for others needing products shipped. 'Private fleets' are subsidiaries of non-trucking firms, and are used almost exclusively to ship their own products. For example, Stop & Shop and

[4]The fact that this segmentation occurs lends power to parts of my empirical work in which firms are grouped into sectors of operation.

Coca-Cola own private fleets that are used to distribute their goods from warehouses to retail outlets. For-hire carriers differentiate themselves to shippers based on several factors. One such dimension is length of haul, which is often classified as local (less than 50 miles), short-range (50 to 200 miles), medium-range (200 to 500 miles), and long-range (over 500 miles). Second, different shipping equipment is often needed for different products. Tank trucks are used to carry hazardous materials, such as chemical and petroleum products, refrigerated vans transport goods requiring cold storage, platform trucks are often used for construction equipment or bulk materials, and dry cargo vans carry products requiring no special treatment. A third segmentation of the industry is based on the size of the shipment, the so-called truckload (TL) and less-than-truckload (LTL) sectors of the industry. The TL sector consists of trucks that carry very large shipments, each haul often stemming from a single shipper. The LTL sector is composed of trucks that combine many small shipments from several sources. The United Parcel Service (UPS) is an example of a carrier that operates in the LTL sector. Finally, carriers offer different terms of the service contract. Common carriage is a spot-market arrangement between a carrier and a shipper, while contract carriage is a longer-term relationship, usually lasting between six months and two years, covering multiple hauls.

The contracting relationship of interest in this paper is not the one that exists between carriers and shippers, but rather is the interaction between the truck driver and his carrier. The great majority of truck drivers, on the order of 90%, are 'company drivers'. A company driver is an employee of a trucking firm who is paid to ship products using the firm's trucks and equipment. The remaining 10% of drivers are 'owner-operators', who own (or lease) their own trucks and are hired by trucking firms on a haul by haul basis. The focus here is company drivers, and in particular how the adoption of the OBC technology causes them to alter their driving behavior.

## 2.2 On-Board Computers, Incentives, and Coordination

### 2.2.1 On-Board Computers

The introduction of OBCs in the late 1980's was recognized as a potentially industry-changing innovation. Trip recorders, introduced slightly earlier than the more sophisticated EVMS, keep a summary of how a driver operates a truck. At the beginning of a haul the trip recorder is activated. Only when the driver returns back to his home base are the contents of the trip recorder accessible, at which time they can be analyzed by the driver's superiors. The information collected by a trip recorder includes departure and arrival times, speed of the truck, revolutions per minute of the engine, idling time, periods of stop-and-go driving, brake use, and precise measures of fuel consumption. Contained in this mass of operating data are the three most important forms of driver behavior that reduce fuel efficiency and shorten engine life: excessive speed, idling time, and over-revving of the engine. The contents of the trip recorder are valuable for mechanics who may need to diagnose engine problems, but more importantly, the data allow firms to observe exactly how drivers drive trucks. This knowledge is of great use to trucking firms since a truck's value, as well as a truck's fuel efficiency, is sensitive to how it is operated. The incentive-enhancing information provided by trip recorders allows firms to better shape driver behavior through more efficient contracting with their employees.[5] A trip recorder cost about $500 to purchase and install in the early 1990's, and the price remained relatively constant through the end of the decade.

EVMS provide carriers with all of the information-collecting capabilities of trip recorders, as well as several additional features. First, they allow real-time communication between drivers on the road and dispatchers at the firms through e-mail type messaging. Com-

[5]This capability of trip recorders (and EVMS) to observe driver actions is referred to as the 'incentive' effect of the technology.

munication without EVMS can be extremely difficult. CB radio can be used, but only when parties are within 25 miles of each other. Cellular phone use by truckers can result in expensive roaming fees, and thus cell phones are rarely used by drivers.[6] Most drivers have to resort to pulling over, stopping their truck, and finding a pay phone in order to communicate with the firm dispatchers. Notice also that without EVMS (or a cell phone) the firm cannot initiate contact with the driver; the dispatcher must wait for the driver to 'check-in' from the road. Thus, the two-way communication feature of EVMS provides a substantial upgrade from the alternative means of establishing contact. Second, through global positioning satellite technology (GPS), EVMS allow firms to track the exact locations of their trucks. Third, the information collected by EVMS is available to dispatchers in real-time, so trucks can be rerouted and schedules reorganized immediately. Finally, the GPS technology allows carriers to provide real-time tracking information to customers on the location of their shipments. Schneider National, one of the largest for-hire carriers and an early adopter of EVMS, cited this feature as a primary reason for installing the system. According to the director of their information services division, "We believed our biggest area of savings would be customer service. It becomes our responsibility to call customers and tell them the status of their shipment, whereas before, they had to call us. The customers don't have to monitor the shipment anymore."[7] The cost of outfitting a fleet with EVMS and integrating their capabilities into firm operations can be substantial. The overall cost fell by about one-third from the early 1990s to the end of the decade. In 1997, a single terminal cost between $2,500 and $4,000, with monthly communication fees of $50 to $100 per vehicle. Additional installation costs can be as high as a few thousand dollars. EVMS' reliability improved greatly during this decade, given the enhanced computing power and

[6]The price of using cellular phones has decreased substantially in the United States in the past two years, but this period is beyond the time frame considered in the empirical work below.

[7]Quoted from Schrodt (1989: 2B).

more sophisticated wireless radio technology that characterized the latter half of this period.

### 2.2.2 The Incentive Effect

In order to better understand the impact of OBC use on driver incentives, it is necessary to provide a more complete picture of a driver's job within the firm. Drivers and carriers operate in the context of a principal-agent relationship.[8] The driver's objective is to maximize his utility, which depends primarily on the income earned and effort expended in completing a shipment. The firm wishes to maximize its profit, which is a function of driver effort.

While the essence of a truck driver's job is to move cargo from one place to another, it is not appropriate to model this job as consisting of a single task. In particular, the *manner* in which cargo is transported by the driver is of great importance to his employer. As a useful abstraction, consider a model where the company driver's role in production consists of two tasks: (i) to transport the product in a timely fashion from one location to another (the 'productivity' task), and (ii) to drive the truck in a manner that is not abusive and maintains the truck's value (the 'operation' task). As noted above, a truck's value is highly dependent on the driver's actions. Poor driving technique (characterized by, for example, driving at high rates of speed, over-revving the engine, idling excessively, accelerating quickly, and shifting erratically) stresses the mechanical structure of the truck, potentially causing part failures and more frequent breakdowns. This behavior also reduces fuel efficiency, which is of interest to the carrier since it is the party that pays for fuel expenses. Finally, high speeds can increase the probability of an accident, which may damage the truck, shipment, and/or driver. The driver, however, may prefer to drive the truck in a way that is not desirable

[8] For a more formal discussion of the driver-carrier relationship in a principal-agent model, see Cacciola (2002).

to the carrier. For example, maintaining a higher average speed while on the road allows the driver to take longer breaks, and yet reach his destination on time. It is this tension between the firm's objectives and the driver's preferences in driving technique, as well as the difficulty in observing the effort directed towards this task, that is at the heart of the contracting decision.

The adoption of an OBC (either a trip recorder or an EVMS) has a clear impact on the observability of the tasks. The productivity task is nearly perfectly observable by the carrier at little cost, even without an OBC. Late arrivals or damaged products are generally reported by the shipper to the carrier. Also, factors outside of the driver's control, such as traffic and weather conditions, are easily verifiable. The amount of effort directed towards the operation task, on the other hand, is extremely difficult to measure without an OBC. Measures of this task are very noisy, in part because carriers cannot easily distinguish between mechanical problems caused by bad driving and those associated with the normal wear and tear of truck use. This is compounded when shipment schedules and truck assignments dictate that more than one driver use the same truck for different hauls. The use of an OBC allows trucking firms to *precisely* monitor driver effort directed towards the operation task. This information can then be explicitly written into driver contracts, or can be used more informally to reward or punish driver behavior. Anecdotally, the incentive effect of OBCs is substantial. A spokesperson for Fleet Boss, a large manufacturer of OBCs, says of driver reaction to the technology: "It's an instant character builder. As soon as employees know what it does, they alter their behavior."[9]

[9]Quoted from Phipps (2001).

### 2.2.3 The Coordination Effect

The key informational difference between trip recorders and EVMS is that trip recorders do not furnish data to dispatchers in real-time, nor allow for immediate communication between drivers and firms. As such, trip recorders can only be used to discern driver behavior and provide for improved incentives. EVMS allow not only for this incentive effect, but the instantaneous data received from trucks and the facile means of communication can improve resource allocation decisions and yield positive 'coordination' effects. Coordination refers to the process whereby dispatchers assign drivers and trucks to shipments. Dispatchers, the individuals based at the firms who arrange the truck and driver schedules, work in an environment where the parameters of their production decisions are constantly evolving. Trucks on the road can be delayed by weather, traffic, and mechanical problems, forcing dispatchers to rearrange schedules, while new shipping orders continue to be placed by customers. At the larger firms, which employ hundreds of trucks and drivers, complex matching algorithms are used to efficiently incorporate new information into truck, shipment, and route assignments. EVMS adds to this flow of information, increasing dispatchers' options and potentially enhancing resource allocation within the firm.[10]

### 2.2.4 The On-Board Computer Adoption Decision

The firm-level decision of whether to invest in OBC technology, and if so, what percentage of the fleet to equip, is ultimately a cost-benefit analysis. The perceived benefits of adoption, measured by improved incentives and/or coordination, are compared with the costs

[10]In my empirical work, I estimate a joint effect of the impact of EVMS on the resource allocation improvements discussed here and the customer service enhancements described above. While it would be ideal to be able to separately identify these two components, the joint effect is certainly of interest since it captures the total gross benefit that trucking firms derive from the vast information capabilities provided by EVMS.

of purchasing, installing, and utilizing the OBC systems. A formal model of this decision, based on profit-maximization, for example, could be developed. But an intuitive discussion of the relevant costs and benefits can lend sufficient insight to the construction of empirical strategies designed to evaluate the effects of OBCs. In terms of the cost of adopting, trip recorders are primarily a per truck expense. There is a small fixed cost in the way of a computer and software needed at the firm to read and process the trip recorder's data, but the majority of the investment is in the purchase of individual trip recorders for trucks. EVMS investment, though, requires a large fixed cost. There is a need for 'back-office' software to optimize routing decisions and truck allocations, and for skilled dispatchers who can use this advanced technology and operate in an information-intensive work environment. This is in addition to the substantial per truck expense required for purchase, installation, and satellite communication fees.

On the benefit side, trip recorders increase productivity on a per truck basis. The marginal productivity of a trip recorder on one truck is independent of whether or not a trip recorder is installed on another truck. With EVMS, though, there are significant spillovers and economies of scale in adoption within a firm. While having EVMS on one truck is of only limited value, outfitting more trucks allows firms to take increasing advantage of the coordination capabilities. As a particular example of the cost-benefit analysis I am describing here, consider the adoption of OBCs across carriers of different sizes. The spillover nature of the EVMS benefit coupled with the large fixed cost of EVMS investment helps explain why larger carriers have much greater adoption rates of EVMS than smaller carriers do. Trip recorder adoption, while somewhat higher for larger carriers than smaller carriers, does not exhibit the sharp discrepancy across fleet sizes as EVMS does.

It is also important to note that the magnitudes of the incentive and coordination benefits vary by sector of the trucking industry. For example, when dispatchers face few

constraints in how they respond to new information, the marginal benefit of coordination is higher than when dispatchers have little use for information improvements. This observation that there is variation in the benefits of incentives and coordination across sectors is the basis of Hubbard (2000), who uses the differential adoption rates of OBCs across sectors to infer their value to firms.

## 2.3 Detecting the Incentive and Coordination Effects in Data

OBCs clearly possess features that can improve driver incentives and firm coordination. Empirically, I see two main questions of interest in evaluating these components of the technology. First, can we detect the presence of these effects in the data? Second, if these effects do exist, what are the magnitudes of the improvements? The Hubbard (2000) approach looks at the variation in adoption rates, and from this variation estimates the value of the incentive and coordination components. The reanalysis in Section 3 below illustrates the fragility of this empirical approach. Namely, there are many factors, aside from incentives and coordination, that influence the OBC adoption decision. For example, I explore the alternative hypothesis that the truck replacement decision drives a significant fraction of the variation in OBC use rates across sectors. Accounting for truck replacement does in fact significantly alter the interpretation of the results. Now, it is possible that the Hubbard (2000) methodology can detect the existence of the incentive and coordination effects, *if all other sources of variation in adoption are accounted for.* But even if this strong condition is satisfied, the framework does not provide a measure of the *size* of the relevant effects. To document that one sector has greater EVMS adoption rates than another may indicate a desire for coordination improvements in the former sector, but it provides no estimate of the extent to which the high adopting sector improves its operations.

A more robust and informative approach is to look at the *impact* of OBC use on *direct*

measures that embody the incentive and coordination improvements. First, consider the detection of the incentive effect. The incentive feature of OBCs allows firms to observe how drivers operate trucks. And if this particular attribute is useful, then drivers should drive differently when they have an OBC terminal installed than when they don't. In which dimensions will drivers alter their behavior? Consider this account of driver conduct offered by a trucker:

> A ... common abuse of machinery centered on driver's attempts to increase output or minimize the amount of time it took to complete a job ... [I]t was not uncommon for drivers at all companies to climb or descend mountain grades at the limits of their trucks' capabilities ... Some drivers secretly altered their trucks' fuel pumps to increase the engine's horsepower, a practice known as 'jacking up.' A jacked-up pump is likely to cost the owner by cutting fuel mileage, lowering the engine's life expectancy, and putting more wear on the drive train and drive tires.[11]

This type of vehicle abuse becomes observable to carriers upon adoption of an OBC, suggesting that the change in driver behavior may be manifested in outcomes associated with vehicle operating characteristics. In particular, I consider as outcome variables the financial *costs* to firms of maintaining their fleets of trucks. The data that I use, described below, contain five relevant cost variables: (i) fuel costs, (ii) outside maintenance costs, (iii) vehicle parts costs, (iv) tires and tubes costs, and (v) total operating supply costs (this fifth variable is the sum of the first four). Driver behavior feeds directly into these costs which are borne by the carriers. To estimate the incentive effect, the empirical implication is that, all else equal, the greater the incidence of OBC adoption within a firm (or a sector), the

[11]Quoted from Ouellet (1994: 85-86).

*lower* the financial costs.[12]

The coordination effect is inherently a scale-enhancing capability, and for this reason I identify it by looking at the impact of EVMS adoption (net of trip recorder adoption) on firm *revenues*.[13] This argument can be justified on two grounds. First, the coupling of the GPS and communication features allows dispatchers to potentially schedule a greater number of shipments. Consider the case in which a dispatcher receives a new order to move cargo from Miami to Boston. The satellite technology allows the dispatcher to know the location and capacity utilization of each truck in the fleet. The dispatcher can then determine which trucks are in the area of Miami, or will be passing through shortly, as well as whether any of the trucks has room in its trailer to hold the new cargo. Once the dispatcher identifies an appropriate truck, the driver can be contacted immediately and instructed as to where to pick up the new shipment. Without the EVMS technology, the dispatcher might not have an accurate assessment of which trucks are in the area, will not be able to initiate contact with any of the drivers, and may in fact have to refuse the new order. EVMS thus reduces inefficiencies in terms of idle firm resources, allowing for a greater number of shipments and an increase in revenues. Second, the ability of EVMS to track shipments (and, with the appropriate software, packages within shipments) is highly valued by customers, particularly those in the retail sector.[14] The impetus towards inventory reduction in several industries means that trucks must run on tight schedules, and shipment tracking is necessary to satisfy customer demands. Offering this service to

[12]Cacciola (2002) studies other measures of the change in driver behavior due to OBC adoption, including fuel *efficiency* (as opposed to the fuel *costs* analyzed here) and truck life expectancy.

[13]By 'scale-enhancing' I do not refer to an improvement in the economies of scale in a firm, but simply mean an increase in the size of a firm's operations.

[14]One survey of 270 carriers in Canada found that customer service enhancement was the number one reason for information technology adoption, followed by improved operations planning, better dispatching capabilities, and improved communication with drivers (see Bigras, Crainic, and Roy (1997)).

customers effectively opens sectors of the industry to the carrier, generating new business opportunities and a boost to firm revenues.[15] Section 4 discusses in more detail the financial data and identification strategy.

# 3 A Re-Analysis of Previous Methods

Hubbard (2000) attempts to estimate the incentive and coordination effects of OBCs by comparing the use rates of trip recorders and EVMS across different sectors of the trucking industry. Given the various characteristics of different segments of the industry, some have a need for improved incentives and others have a need for coordination enhancing instruments. With the introduction of OBCs, trucks and firms that have a large scope for the incentive effect can benefit from using trip recorders, and trucks and firms that have a large scope for coordination improvements (relative to the incentive benefits) will find more value in EVMS relative to trip recorders. Assuming that the OBC adoption decision is part of firms' profit maximization, it follows that firms with high adoption rates of trip recorders benefit from the incentive effect, and that firms with high adoption rates of EVMS (relative to trip recorders) benefit from the coordination effect.

## 3.1 Testable Propositions

Based on Hubbard's observation about the relative benefits of trip recorders and EVMS, he establishes several empirically testable propositions about on-board computer adoption. Given their incentive effect, trip recorders should be used *more* frequently in the following

[15]There is also the possibility that the *incentive* effect of OBCs can enhance revenues. For example, this might be the case when drivers spend a significant amount of time on non-driving activities, such as loading and unloading the truck. The amount of time spent performing these duties can be monitored with an OBC, alerting the firm to shirking by drivers. This scenario is discussed more below.

cases:

1. Trucks stop infrequently. For example, **long-haul trucks** that spend several days away from their home base afford their drivers more latitude in how the truck is operated, and thus we would expect to see a high incidence of trip recorder use. Likewise, conditional on length of haul, trucks operating in the **TL sector** should have a greater use rate than trucks in the LTL sector. LTL trucks make frequent stops throughout the day, often running regular routes and returning to their base terminal at the end of a shift.

2. Late arrivals are costly. This is true of trucks that deliver to loading docks, where drop-offs tend to be precisely scheduled. Late arriving trucks leave resources underutilized and potentially delay other arriving shipments from unloading. Use of a trip recorder will allow the firm to ascertain the cause of the driver's tardiness. An example here is **refrigerated trucks**.

3. Accidents are particularly costly. A record of a driver's actions can be valuable for a firm when settling claims with insurance companies. Accidents are relatively more costly when trucks carry hazardous cargo, such as chemicals and petroleum. The trucks used to carry this type of cargo are **tank trucks**.

The propositions involving EVMS use relative to trip recorder use (and thus identifying the coordination effect) are the following:

1. Relative EVMS use should be higher when dispatchers face few constraints in how they respond to new information and orders. Private fleets often have many of their trucks occupied in regularly scheduled in-house shipments, leaving little scope for dispatchers to reallocate trucks. In this case the coordination-related information provided by EVMS is only of limited value. Dispatchers in for-hire firms, on the other

hand, are allowed more discretion as new orders arrive. Within for-hire firms, the spot-market arrangements of common carriage are less constraining than the longer-term commitments in contract carriage. Overall then, EVMS use should be greatest in **common carriage**, followed by **contract carriage**, and least in **private fleets**.

2. Relative EVMS use should be *low* when OBCs are only used for verification purposes. Since knowledge of driver behavior is sufficient for verification, the coordination component of EVMS is not needed, and EVMS adoption will be low relative to trip recorder adoption. This is true of **tank trucks**, which haul petroleum and chemicals.

## 3.2 Data

The data used in this section is from the Census of Transportation's Truck Inventory and Use Survey (TIUS) for the years 1987, 1992, and 1997.[16] The TIUS is a random sample of the United States' truck population, providing data on physical and operating characteristics. This paper uses the observations on truck-tractors, which are the front-end power-units of the truck-and-trailer combinations. For the 1987, 1992, and 1997 sample, there are 24,989; 39,850; and 25,533 observations (trucks), respectively.[17] Key variables included in the TIUS are trip recorder and EVMS use, length of haul, type of trailer attached, principal products hauled, fleet size, model year, and operation in a for-hire or private fleet. For the trucks in the for-hire sector, there are further breakdowns into TL and LTL, as well as common and contract carriage. The surveys do not form a panel, and there is no way to match individual trucks across years or to identify trucks operating in the same firm.

[16]The name of the survey changed to the Vehicle Inventory and Use Survey in 1997.

[17]Due to missing variables, some specifications use fewer than the maximum number of observations.

### 3.3 Empirical Results

Tables 1 through 3 provide summaries of trip and EVMS adoption in particular sectors of the industry.[18] In 1987 neither trip recorders nor EVMS were available, so adoption begins to emerge in the 1992 survey. In all sectors in 1992, about 7.5% of trucks were equipped with trip recorders, and roughly 10.5% had EVMS installed. By 1997, trip recorder use edged up to 8%, while EVMS adoption exploded to nearly 25% of trucks. Table 1 looks at the use rates by average length of haul. There is considerable variation across categories in both trip recorder and EVMS use. There is a nearly monotonic increase in trip recorder adoption as average length of haul grows, consistent with Hubbard's prediction. Interestingly, the same pattern is true of EVMS adoption. The robustness of this latter phenomenon is tested and discussed below. Table 2 summarizes OBC use by the type of trailer most commonly pulled by the truck-tractor unit. Tank trucks have a high rate of trip recorder adoption, consistent with the proposition above, and refrigerated vans have high use rates for both trip recorders and EVMS. TL, LTL, and private fleets are summarized in Table 3. The TL sector has greater trip recorder adoption rates than the LTL sector, while private fleets have below average EVMS use rates; both facts are consistent with the empirical propositions.

Table 4 is indicative of Hubbard's (2000) approach to more formally testing the empirical propositions above. This table uses the 1992 TIUS, a sample identical to that used in Hubbard (2000). Here, two separate linear probability models are estimated. In the first, the dependent variable is OBC use, which is the sum of trip recorder and EVMS use.

[18] In these tables, as well as in all of the regressions that follow in this section, observations are weighted by the expansion factors provided by the Census. The trucks sampled in the TIUS were selected using stratified randomization, where the strata were defined by state and five truck-type categories. Since I use only one of these truck-type categories, truck-tractors, the use of expansion factors corrects for the over-sampling of smaller states. Replication of the analysis without weighting by the expansion factors yields nearly identical estimates to the weighted version presented here.

The second regression contains only trucks with an OBC installed, and models the EVMS versus trip recorder adoption decision. The first regression will identify a joint effect of incentives and coordination. The second regression, since it estimates EVMS use relative to trip recorder use, will isolate the pure coordination effect, assuming that the regression is correctly specified. The covariates in both equations include the variables of interest (length of haul, trailer type, TL, and private fleet/contract type) as well as controls for intrastate operation, whether the truck is driven by an owner-operator, if the truck refuels at a private facility (as opposed to a truck stop), exempt carrier status (a vestige of regulation), fleet size, principal product hauled, and base state of operation. Looking first at the estimates for the trucks that have an OBC installed, the propositions for the coordination effect are all supported. Both private fleets and trucks operating under contract carriage are less likely to use EVMS for coordination purposes than common carriage, which is the omitted category (20% and 6% less likely than common carriage, respectively, all else equal). The tank truck coefficient is negative and significant (a point estimate of -.177), also in accordance with one of the propositions. Though not predicted by the theory, the EVMS use rates for trucks operating over 200 miles from home are quite large and significant (compared to trucks operating locally, 13% and 15% more likely for the dummy variables for 200 to 500 miles and over 500 miles, respectively), as is the TL variable (a point estimate of .070).

As indicated above, the first column jointly estimates the incentive and coordination effects. The pure incentive effect of the technology can be isolated by a comparison of the estimates from the first and second columns. Simply take the desired coefficient from the OBC/non-OBC regression and subtract off the coefficient in the EVMS/trip recorder regression times $FRAC$, where $FRAC$ is the number of trucks with an EVMS divided by the number of trucks with either type of OBC. The adjustment by $FRAC$ weights the coordination estimate identified in the second column by the incidence of EVMS use.

Intuitively, if EVMS use is high relative to trip recorder use, then a large component of the first column estimate is comprised of coordination. The formula adjusts for this by purging a greater share of the second column estimate from the first column estimate, leaving the pure incentive effect. As an example, consider the incentive effect for the TL variable. Since $FRAC$ is .55, (3.311/6.023), the incentive effect is $.137 - (.070)(.55) = .099$, indicating that trucks in the TL sector adopt OBCs for incentive purposes at a rate 9.9% greater than trucks in the LTL sector. This is consistent with one of the propositions. Similar computations of the incentive effect along other dimensions (length of haul, tank truck, and refrigerated van) also yield results consistent with the empirically testable propositions.[19]

### 3.4 The Consideration of Truck Age and Replacement

Another dimension of potential importance in OBC adoption rates is the model year, or equivalently, age, of the truck. Figures 1 and 2 are a first step in considering this relationship. These figures are bar graphs which illustrate trip recorder and EVMS use rates broken out by model year, with Figure 1 using the 1992 survey and Figure 2 using the 1997 survey. Several features are noteworthy here. First, OBC use increases virtually monotonically in model year in both surveys. Second, this increase is almost exclusively the result of EVMS adoption, as trip recorder use is fairly constant across model years at just under 10%; moreover, this is true for both the 1992 and 1997 surveys.

In considering the implications of these figures for tests of the empirical propositions, note that Table 4 uses both the within and between model year variation in OBC adoption rates. Given that EVMS adoption is trending upward across model years and trip recorder use is constant, the framework implicitly supporting Table 4 would attribute these differen-

[19]The incentive effect estimates for the four length of haul variables, from shortest to longest, are .021, .077, .072, and .066, for the tank truck variable is .125, and for the refrigerated van is .071.

tial trends to an increased desire for the coordination-enhancing benefits of EVMS. This is a possibility, but other explanations for the figures are equally likely. Clearly the decreased price and increased reliability of EVMS later in the decade made installing EVMS on new trucks more attractive, independent of any desire for the coordination benefits. This may present difficulties in interpreting the results of Table 4, depending on the frequencies of new truck purchases by differing sectors of the industry. If the sectors predicted to benefit by OBCs are also replacing their trucks more frequently, independent of their desire for OBCs, then the coefficients in Table 4 may spuriously indicate coordination and incentive effects where none actually exist.

The solution to this potential problem is simple. Instead of using both the within and between model year variation, we can limit ourselves to looking within model years. One way achieve this is to break up the analysis by model year (age). As an example, Table 5 displays the results for the TL versus LTL comparison. Recall that in Table 4, the coefficients indicate both an incentive and coordination effect in the TL sector. Table 5 uses a slightly different specification than Table 4; it is the same in spirit and only differs in interpretation. The first and second columns are linear probability models for trip recorder use and EVMS use, respectively. The rows represent different model years.[20] Note that each cell in the table is the TL coefficient from a regression using only trucks of the relevant model year. The results are fairly striking. For model years 1985 and older, trip recorder use in the TL sector is significantly above that in the LTL sector (with the exception of the pre-1982 trucks, where there is no difference), and EVMS use is not significantly different in the two sectors. But for trucks made after 1985, *trip recorder* use is no different between the two sectors, and *EVMS* adoption is significantly higher in the TL sector. Also, EVMS

[20]The TIUS survey includes 11 age categories: new, separate categories for one through nine years old, and one group for trucks ten years or older.

adoption appears to be trending upward in the TL sector as compared with the LTL sector for newer trucks. As mentioned above, it seems unlikely that the coordination benefits of being in the TL sector are substantially changing over time: it's more likely that EVMS are being purchased partly in place of trip recorders as firms buy new trucks.

The reason that I use this particular specification in Table 5, rather than the one used in Table 4, is to illustrate the 'switching point' from trip recorders to EVMS in model year 1985 and the trend in EVMS adoption across model years. This stark trend is very compelling evidence that truck replacement and diffusion of the technology are integral parts of the EVMS versus trip recorder adoption decision. This argument is confirmed by breaking out the analysis by model year using the specification in Table 4. For trucks which have an OBC installed, there is *no* significant difference in EVMS adoption between the TL and LTL sector for *any* of the model years. But there are significant differences (at the 5% level) in OBC adoption and non-OBC adoption between the TL and LTL sectors in eight of the eleven model years. This is consistent with an incentive effect in the TL sector but no coordination effect, indicating a spurious finding of a coordination effect in Table 4.

To explore the truck replacement effect more generally, Table 6 analyzes the purchasing patterns of trucks across different sectors of the industry. Looking first at the results for the 1992 survey (in the second column) we see that the sectors with younger trucks are precisely those that have higher OBC use in Table 4. Truck age decreases with length of haul, is lower for tank trucks, refrigerated vans, and for trucks in the TL sector. Clearly, Table 4, since it does not consider the age variation across sectors, is confounding the monitoring effects of OBCs with the fact that newer trucks are more likely to install an EVMS relative to a trip recorder. There is also a possibility that the sectors buying new trucks are perhaps replacing trucks with the idea that newer trucks are more compatible with the EVMS technology. In other words, the OBC use decision and the truck replacement decision are

jointly determined by a firm. But this behavior can be ruled out using the 1987 survey, since it was conducted prior to the introduction of OBCs. Firms surveyed at this time will therefore purchase new trucks independent of any OBC considerations. If the results for the 1987 and 1992 surveys are similar, this is an indication that sectors are not changing their purchase patterns because of the new technology. This is precisely what we see in Table 6. With very few exceptions, the results from the 1987 survey are nearly identical in sign, and often in magnitude, to the results from the 1992 survey.

The preceding discussion is meant to support the argument that the between model year variation is corrupted by the fact that different sectors of the trucking industry replace trucks at differential rates. And since newer trucks are more likely to use EVMS relative to trip recorders, *independent* of incentive and coordination effects, we may spuriously infer the existence of monitoring effects. The proposed solution is to use only the within model year variation. The simplest estimation strategy that accomplishes this is one that augments the specification in Table 4 to include a set of truck age dummy variables as covariates. A joint test of the significance of these age dummies indicates that they drive a considerable portion of the variation in OBC adoption ($F(10, 34975) = 58.77$ yielding a p-value of 0.0000 in the OBC/non-OBC regression and $F(10, 5963) = 49.01$ yielding a p-value of 0.0000 in the EVMS/trip recorder regression). The problem with this simple fix is that it masks substantial heterogeneity in the sector coefficients for different model years. A more suitable specification allows the sector coefficients to vary, a strategy that I employ by estimating the regressions in Table 4 separately for each model year. I performed this exercise for the 11 age categories, and then, due to power considerations and a desire for parsimony in the presentation of the results, decided that allowing for three age categories (new trucks, trucks ages 1 to 4, and trucks older than 5 years) is sufficiently flexible to capture the heterogeneity in the coefficients. I then re-estimated the model separately for

each of these three age categories.

Tables 7 through 9 present these regressions. First, it should be noted that the restriction that the sector coefficients are equal across model years, which is imposed in the pooled model year regression with truck age dummies, is rejected when compared with a model where the coefficients on length of haul, trailer type, TL, and private fleet/contract carriage are allowed to vary across the three age groups ($F(22, 34975) = 8.82$ yielding a p-value of 0.0000 in the OBC/non-OBC regression and $F(22, 5963) = 4.01$ yielding a p-value of 0.0000 in the EVMS/trip recorder regression). In the EVMS/trip recorder regressions, there is quite a bit of heterogeneity in the point estimates for the tank truck variable, as well as the TL variable. The tank truck variable coefficient is -.072 for new trucks, falls to -.227 for trucks ages 1 to 4, and is -.111 for trucks older than 5 years. While all of these estimates are within sampling error, the differences between them cloud inference regarding the coordination effect. The TL coefficients are .005, .084, and -.069, respectively, for the new trucks, middle aged trucks, and the oldest trucks, a fair amount of variation which is masked in the pooled model year specification.

Taken individually, the within age group regressions do not yield as strong or as clear a set of findings of a coordination effect as does Table 4. In Table 7, which uses new trucks (model year 1992), the tank truck coefficient is no longer significant in the EVMS/trip recorder regression. Notice also, that while not included in the empirical propositions, the TL and length of haul variables are not significant, as they were in Table 4. The private fleet and contract carriage coefficients, though, are significantly negative as predicted by the theory. The results in Table 8, for trucks ages 1 to 4, are similar, though the contract carriage coefficient becomes insignificant while the tank truck coefficient is significantly negative, as predicted in the propositions. In Table 9, the sample of trucks over 5 years old, for the coordination effect only private fleet is significant. Certainly some of the statistically

insignificant results in the within age group regressions are the result of lower power due to a reduction in sample sizes from Table 4. But the disparities in the coefficients across age groups for several of the sector variables illustrate the tenuous nature of this estimation scheme.

The incentive effect estimates in Tables 7 through 9 remain consistent with the empirically testable propositions. Using the simple formula described earlier, the implied incentive effect for length of haul, tank truck, refrigerated van, and TL is positive for each variable compared with its omitted category. The reason that the incentive effect is robust to the within age group correction is clear: truck replacement affects adoption primarily along the EVMS/trip recorder margin, as opposed to the OBC vs. non-OBC decision. This is best illustrated in Figures 1 and 2, where EVMS use is shown to grow dramatically over time compared to trip recorder use. It is also important to emphasize that though the within age group regressions tend to diminish the findings of a coordination effect, this does not mean that the EVMS technology does not lend itself to coordination improvements. Rather, the within age group analysis highlights a flaw in the methodology, namely that other omitted factors, aside from truck age, may also drive the OBC adoption decision. Using the variation in OBC adoption *in isolation* is a dangerous way of identifying the coordination, and even the incentive, effect of OBCs. To credibly estimate both of these effects it is necessary to look at the impact of EVMS and trip recorders on relevant outcome variables.

# 4 The Impact of On-Board Computer Adoption on Costs and Revenue

As an alternative to using the variation in adoption rates itself to identify the value of OBCs, this section examines the more direct means of projecting this variation onto measures of

the incentive and coordination capabilities. The impact of OBCs on firm financial costs (via changes in driver behavior) identifies the incentive effect, and the effect of EVMS use (relative to trip recorder use) on firm revenue identifies the coordination effect. To empirically implement this idea, I create cell-level data from OBC adoption data and firm-level finances. Both of these data sources provide observations prior to, and of course after, the introduction of the OBC technology to the trucking market.

## 4.1 Data

One source of data used in this section is the TIUS collected by the Census, described earlier, which includes the OBC adoption information. Here, I use the 1987 and 1997 surveys.[21] The other source of data is firm-level financial data collected on for-hire motor carriers. The largest carriers, classified as Class I and Class II, are required to file detailed annual reports with the Bureau of Transportation Statistics (BTS).[22] Carriers are required to file this 'Form M' information, unless they request and are granted a confidentiality exemption. This exemption is not trivial to obtain, as carriers must at a minimum demonstrate that public release of the data causes competitive harm and that there is a need to preserve confidential business information. The BTS provides for public use all of the data for nonexempt carriers; this data is called the Motor Carrier Financial & Operating Statistics Annual Reports (MCFOS). The data contain balance sheet (assets and liabilities) and income statement (operating revenues and costs) reports, as well as information on the type of service offered, tonnage, mileage, employees, and transportation equipment. I use data from the 1989 and

[21] I limit my use of the TIUS to observations on trucks operated by company drivers working in for-hire fleets in order to more precisely match the sample of firms surveyed in the financial data. In the 1987 and 1997 surveys, this restriction leaves 36% and 41% of the original truck-tractor observations, respectively.

[22] In 1999, carriers were categorized as Class I if annual operating revenue exceeded $10 million, and were Class II if annual operating revenue was between $3 million and $10 million.

1999 surveys, which include data on 1,714 and 1,706 carriers, respectively. Unfortunately, the years of the TIUS and MCFOS do not correspond exactly. This is because the TIUS data is only collected every 5 years, and the MCFOS is very difficult to find.[23] As such, the 1987 TIUS data is matched with the 1989 MCFOS data, and the 1997 TIUS is matched with the 1999 MCFOS.

## 4.2 Empirical Framework

The firm-level relationship that I am interested in estimating and using to identify the incentive and coordination effects is

$$Y_{it} = \beta EVMS_{it} + \gamma TR_{it} + SECTOR'_{it}\delta + X'_{it}\psi + \alpha_i + \epsilon_{it} \quad (1)$$

where $i$ indexes firms and $t$ indexes time; $Y$ is the dependent variable, which is some form of costs or revenue; $EVMS$ and $TR$ are the fraction of truck-tractors in a firm with EVMS and trip recorders installed, respectively; $SECTOR$ is a vector of dummy variables indicating the type of service provided by the carrier (this includes length of haul, cargo type, and base state of operation); $X$ is a vector of additional control variables, including the truck age distribution and fleet size of a firm; $\alpha_i$ is a firm-level fixed effect, which includes such factors as managerial ability and technological savvy; $\epsilon_{it}$ is a white-noise error; and $\beta$, $\gamma$, $\delta$, and $\psi$ are parameters. The incentive effect, which is estimated where the dependent variable is some form of costs, is identified by $\beta$ or $\gamma$ (which are hypothesized to be negative). EVMS and trip recorders provide identical information on driver behavior, so should yield the same effect on costs. The implication is that $\beta$ and $\gamma$ are equal, which is easily testable. The coordination effect is estimated in a regression where revenue is the

[23]The BTS itself only maintains the MCFOS as far back as 1996, so earlier data must be obtained by private firms that specialize in transportation data. The 1989 data that I use is maintained by a firm called Transportation Technical Services, and the 1999 data is from the American Trucking Association.

dependent variable. Since only EVMS has coordination capabilities, while both EVMS and trip recorders provide the means to improve incentives, it is the difference between $\beta$ and $\gamma$ that identifies the coordination improvements, with this difference hypothesized to be positive.[24]

The practical problem in estimating equation (1) is that I don't have firm-level data that includes OBC use information. But I do of course have OBC adoption in a survey of trucks. So in order to combine these two sources of information, I group firms (from the MCFOS) and trucks (from the TIUS) into industry segments. I then compute cell-level averages of the appropriate variables, and use the cross sector and time series variation to examine the impact of OBC use on costs and revenue. In terms of defining the cells, there is a trade-off involved in determining the optimal number of variables used to construct the groups. On the one hand, it is desirable to use a large number of defining variables in order to make the trucks and firms within cells as similar as possible. This has the virtue of preserving a high degree of variation in the cell-level data, as well as creating a large number of cells. But as the number of defining variables increases (and the number of trucks and firms per cell decreases), the cell-level *sample* averages become less precise measures of the true *population* means (this is discussed more below). This induces greater sampling error in the constructed cell-level data.

The cell-level groups that I use are defined by length of haul, cargo type, and base state of operation.[25] This particular grouping provides considerable meaningful variation

[24]The coordination and incentive effects should not be construed as applying purely to revenues and costs, respectively. As alluded to earlier, it may be the case that trip recorders and their corresponding incentive effect have a positive impact on revenue (i.e. in the revenue regression, $\gamma$ is positive). Even if this is true, $\beta - \gamma$ still identifies the coordination effect since $\beta$ captures both the incentive and coordination effects while $\gamma$ reflects only the incentive effect. Similarly, EVMS may lower costs via the coordination effect. The empirically testable form of this hypothesis is $\beta < \gamma$ in the cost regression.

[25]I also replicated the analysis using cells defined by length of haul, cargo type, and fleet size. The results

in OBC adoption rates. Consider length of haul. Since trip recorder adoption is driven by the availability of alternative means of monitoring, trucks that operate close to a home base have low adoption rates, while long-haul trucks have high adoption rates. EVMS adoption is in part a function of how well firms can interact with drivers if a computer is not installed. Short-haul trucks can often use CB radio to communicate with dispatchers, while dispatchers have great difficulty in contacting long-haul truckers without EVMS. In terms of cargo type, trip recorder adoption varies by the use of loading docks (refrigerated trucks) and the cost of accidents (tank trucks), as discussed in Section 3. EVMS adoption is partially a function of the time-sensitivity of hauls. For example, refrigerated vans, carrying products that have a short shelf-life, have a much greater use rate of EVMS than vans carrying dry products. Finally, there is substantial variation in the use of OBCs across states. Trip recorder variation in part reflects geography: states with a mountainous terrain, in which driver behavior has a larger impact on truck value (for example, Utah and Tennessee) have high adoption rates, while flatter states (like Indiana and Ohio) have low adoption rates. Variation in gas prices across states also determines the benefit of monitoring and therefore influences trip recorder adoption. EVMS adoption, since it varies with the availability of other means of communication, tends to be low in states that are densely populated and truck stops are widespread (so truckers can more easily call dispatchers), for example in New England, while adoption is higher in more sparsely populated states, like the Western desert states. This length of haul, cargo type, base state grouping results in between 270 and 321 cells (depending on the specification), on average 5 firms per cell, and 14 trucks per cell.

To derive a form of equation (1) that I am able to estimate, I aggregate over those firms $i$ that belong to cell $c$ that are observed in the data at time $t$. This yields a relationship of

from this grouping are less precise than the ones that I display, though qualitatively similar.

observed sample cell-level averages of

$$\overline{Y}_{ct} = \beta\overline{EVMS}_{ct} + \gamma\overline{TR}_{ct} + SECTOR'_{ct}\delta + \overline{X}'_{ct}\psi + \overline{\alpha}_{ct} + \overline{\epsilon}_{ct}. \quad (2)$$

As Deaton (1985) discusses, there are two main issues involved in estimating this relationship. First, $\overline{\alpha}_{ct}$, the average of the firm-level fixed effects for those firms in cell $c$ that are sampled in the survey, is *not* constant over time, as the *population* mean fixed effect is, since different firms are surveyed in different years. Assuming that $\overline{\alpha}_{ct}$ is correlated with the independent variables, this implies that a fixed effect estimator will not deliver consistent estimates of the parameters. Second, the cell-level sample averages of the variables are error-ridden proxies of the population means. This measurement error in the independent variables will tend to push the OLS and fixed effects estimates of the parameters towards zero. The traditional solution to these estimating problems is to use an instrumental variables strategy. The regressors of interest, here $EVMS$ and $TR$, would be instrumented by variables which are correlated with these regressors but independent of the cell-level unobserved fixed effects. Unfortunately, given the nature of the data that I have, in particular the small sample sizes, there are no good candidates for instrumental variables that provide sufficient explanatory power in the first stage. Instead, I report OLS and first-difference estimates, and interpret them with the above caveats in mind.

It is also important to note that the parameters that I estimate reflect the returns to adoption for those trucks and firms that do in fact decide to adopt the technology. Clearly, the adopters do so for a reason; they can benefit positively from OBCs. Non-adopters, on the other hand, don't adopt because it doesn't behoove them to, perhaps because they operate in a sector of the industry where incentive and coordination issues are negligible or because there are other devices that provide the same features at lower cost (for example, CB radios for short-haul trucks). So the average return for installing an OBC on a random

truck is lower than the estimates that are presented below. The estimates that I provide are certainly of interest though, because they measure the overall impact of OBCs on the trucking industry.

### 4.3 Costs

The impact of OBC adoption on costs is displayed in Tables 10 through 15. In each of these tables, the metric used is the *natural log* of the particular cost variable *per ton-mile*. I use a log specification since it results in a more normal distribution of the dependent variables. A ton-mile is simply one ton of product transported one mile. So, for example, a 10 ton load moved 300 miles would be a 3,000 ton-mile shipment. I normalize costs by ton-miles because, according to industry experts, it is the interaction of tonnage and miles driven that reduces fuel efficiency and causes wear-and-tear on the truck.

In each of the tables, I provide four regressions which vary along two dimensions: weighted or unweighted, and inclusion or exclusion of main effects for the cells (this is the *SECTOR* variable in the above equations). In the OLS regressions, the weighting function used is $n_{99} + n_{97}$ where $n_{99}$ is the number of firms in a given cell in the MCFOS in 1999, and $n_{97}$ is the number of trucks in a given cell in the TIUS in 1997; in the first-difference regressions the weighting function is $n_{89} + n_{99} + n_{97}$ where $n_{99}$ and $n_{97}$ are as already defined and $n_{89}$ is the number of firms in a given cell in the MCFOS in 1989.[26] The weights for $n_{89}$ and $n_{87}$ (the number of trucks in a given cell in the TIUS in 1987) are not included in the OLS regressions because there is no variation in OBC use in 1987. Therefore, the OLS regressions only include the matched cross-section from the late 1990's

[26] Estimates using other weighting functions, such as $\sqrt{n_{99}n_{97}}$ and $\sqrt{n_{99}} + \sqrt{n_{97}}$ for the OLS regressions, and $\sqrt{n_{89}n_{99}n_{97}}$ and $\sqrt{n_{89}} - \sqrt{n_{99}} + \sqrt{n_{97}}$ for the first-difference specifications, were also computed, and do not differ much from the reported weighting schemes.

data. Note also that $n_{87}$ is implicitly set to zero in the first-difference weights since OBC use in 1987 is zero for all cells. There are a couple of reasons why weighting might be preferred. First, putting more weight on cells with greater numbers of firms and trucks helps to deliver the *population* level estimates of the parameters. Second, weighting is an ad hoc fix-up for the sampling error in the regressors. More weight is placed on those cells with a greater number of observations, and it is these cells that provide more precise estimates of their population-level means. The inclusion or exclusion of the main effects for the cells (i.e. dummy variables for length of haul, base state, and cargo type) is really a consideration of power. Starting with only 300 observations, the inclusion of main effects for the cells uses roughly 50 more degrees of freedom. So by excluding them, more of the variation is retained to estimate the OBC adoption coefficients. But of course excluding them when they should be included can bias the coefficients of interest. In fact, the sectors were created in part to provide variation in OBC adoption rates, so omitted variable bias will be present if the dependent variable is significantly related to the main effects of the cells. This issue is easily addressed by testing the joint significance of these main effects. In the majority of the cases they do add significant explanatory power to the regressions. In all regressions I also include as controls the fraction of trucks in each age category (new, one-year old, etc.) and fleet size.

Table 10 presents the cross section OLS regressions where log fuel costs per ton-mile is the dependent variable. F-tests of the joint significance of the cell dummy variables overwhelmingly reject a zero effect in both the unweighted and weighted regressions, indicating that these variables should be included. Overall, the estimates on the OBC variables are very imprecise. In the regressions which include the cell effects, the point estimates for both EVMS and trip recorder are negative, though their imprecision makes inference difficult. The first-difference estimates for log fuel costs per ton-mile, displayed in Table 11, are even

more imprecisely estimated, perhaps a manifestation of measurement error in the regressors.[27] Again, it appears that the main effects for the cells should be included (though there is a borderline significant joint effect, p-value of .074, for the unweighted regressions). Both of the OBC coefficients in the unweighted specification are negative, as is the trip recorder coefficient in the weighted regression, though none of these are close to being statistically significant at conventional levels. The difficulty in rejecting a zero effect of OBCs on fuel costs is perhaps not surprising considering results from other work. Cacciola (2002) finds roughly a 3% improvement in fuel efficiency (miles per gallon) upon adoption of an OBC using a large survey of individual trucks. There is simply not enough power in the matched financial and OBC data to detect an effect this 'small', though it is of nontrivial economic significance to firms.

There is considerably more action when looking at the effect of OBCs on the log of outside maintenance costs per ton-mile. Larger carriers typically do all of their maintenance in-house, using their own mechanics and facilities. Smaller carriers typically employ a mechanic or two to perform routine maintenance and fix minor repairs, but major overhauls tend to be contracted out to vendors. It is these payments to entities not employed by the carrier that is captured in outside maintenance costs. The unweighted cross section regressions in Table 12 show a substantial negative correlation between OBC adoption rates and outside maintenance costs. When the cell effects are not included both the EVMS and trip recorder coefficients are significant at the 5% level. In the more appropriate case where main effects for the cells are included, both point estimates are negative, the trip recorder estimate significantly so at the 10% level. The weighted estimates show less of an impact of

[27]Note that OBC adoption in 1987 is zero, so that the growth in OBC use between 1987 and 1997 is identical to the level of OBC use in 1997. This also explains why there are equal numbers of observations in the cross section OLS and first difference estimates; there is no variation in OBC use in 1987 to identify an effect.

OBCs. All OBC coefficients are negative, but each of the point estimates is smaller than the corresponding unweighted estimate, and none of the weighted estimates is positive. The same qualitative facts are true of the first-difference estimates shown in Table 13. The unweighted estimates indicate a strong negative association between OBC use and outside maintenance costs, while the weighted estimates are smaller and not statistically significant. Since the weighting scheme puts more emphasis on the cells with more firms and trucks, this result indicates that OBCs have more of an impact in the 'niche' markets where only a handful of firms operates.[28] There are two plausible explanations for the maintenance cost results. First, it could be documentation of the incentive effect. Drivers with OBCs are altering their behavior in a way that is less stressful on the truck, resulting in fewer catastrophic engine problems over a truck's lifetime and thus lower payments to outside maintenance facilities. A second possibility is that in-house mechanics are able to deal with a greater range of mechanical problems if the trucks are outfitted with OBCs because of the enhanced diagnostic capabilities of the technology. Empirically, I can't distinguish between these two possibilities, but in practice I think that each plays a part in the large impact of OBC adoption on outside maintenance costs.

The final cost variable that I consider is the log of total operating supply costs per ton-mile. Operating supply costs consist of fuel costs, outside maintenance costs, vehicle parts costs, and tires and tubes costs. The analyses of the latter two costs, not displayed here, show results that are similar in character to the fuel cost tables. They display a pattern of predominantly negative point estimates for EVMS and trip recorders, but lack a precision that would allow for a rejection of the null hypothesis of a zero effect. The same is true of

[28]Explanations for why the attenuation occurs in the weighted estimates are purely speculative. The fact that the weights themselves measure the numbers of firms and trucks in a given cell indicates that the attenuation may be related to the degree of competition within sectors of the industry.

total operating supply costs, displayed in Tables 14 and 15 for the cross section and first-differences, respectively. In the unweighted cross section regressions, both including and excluding the cell main effects, the point estimates are negative but not nearly significant at standard levels. In the weighted estimates, three of the four coefficients across the two regressions are negative, though all t-stats are less than one. The first-difference estimates exhibit a similar imprecision. Seven of the eight coefficients on OBC variables have t-stats less than one, though the point estimates imply fairly large effects. Overall, the cost analysis is troubled by a lack of power. The exception is the analysis of outside maintenance costs, where the results are consistent with a substantial impact of OBC use on driver behavior.

## 4.4 Revenue

Tables 16 and 17 report the results of the relationship between OBC adoption and revenue. Unlike in the cost analysis, I do not normalize revenue by ton-miles. Since the coordination capability of EVMS is scale-enhancing in nature, the benefits may in part be manifested in an increase in total ton-miles. In order to capture all improvements in the scale of operations, I use the log of gross revenue as the dependent variable. The set-up of the revenue tables is the same as the cost tables. Weighted and unweighted estimates are provided, where the same weighting scheme is employed as in the cost section, and I include specifications with and without main effects for the cells. Table 16 presents the results from the cross section regressions of log revenue on EVMS and trip recorder adoption, as well controls for truck age and fleet size. Tests of the joint significance of the cell main effects indicate that these variables should be included in both the weighted and unweighted specifications, so I will focus on these results. The unweighted regressions show a significant association between both EVMS and trip recorder adoption and revenues. The point estimates on EVMS and trip recorder are nearly identical, and indicate that sectors with complete adoption have

revenues on the order of 50% greater than sectors with zero OBC adoption. The magnitude of the trip recorder estimate is somewhat surprising. The coordination effect is identified by the difference between the EVMS coefficient and the trip recorder coefficient, and the discrepancy between the two in the unweighted regressions is basically zero. The weighted cross-section results tell a bit of a different story. The EVMS coefficient is somewhat lower, at .285, but remains significant at the 5% level. The trip recorder coefficient, though, is insignificant, and actually has a small negative coefficient. Taking the point estimates at face value, this discrepancy between the EVMS and trip recorder coefficients is consistent with a coordination effect (note, though, that the point estimates are still within sampling error of each other). Of course in a cross-section, the large co-movement between EVMS and revenue is purely an association, and is not indicative of causality. It could simply be the case that the richer sectors with greater operating revenues are able to invest in more sophisticated technology (if, for example, cash-poor firms face liquidity constraints), and so there exists a positive relationship between EVMS and revenue.

A partial solution to this endogeneity in OBC adoption is to estimate a first-difference model. First-difference estimates will be robust to any time-invariant characteristics of *cells* which are correlated with OBC use. But recall from the earlier discussion that any *firm-level* fixed effects will not necessarily be differenced out, since different firms are surveyed in each year. At the very least, though, first-differencing will purge the estimates of the influence of a portion of the firm-level unobserved characteristics. Table 17 reports the results from this specification where the change in log revenue is related to the change in OBC use (or, identically, the level of OBC use in 1997). The estimates display a similar imprecision as the first-difference cost estimates do, with each standard error in Table 17 greater than the corresponding one in Table 16. Again, it appears that the main effects for the cells are jointly significant, so I will focus the discussion on the unweighted and weighted versions

of these estimates. In the unweighted specification, the EVMS coefficient of .444 is on par with the cross-section estimate, though it is no longer statistically significant. The trip recorder estimate, .062, is lower than in the cross-section, and has a fairly large standard error (.503). The difference between the EVMS and trip recorder coefficients points to a coordination effect, though the estimates are well within sampling error. The weighted estimates show economically large, though statistically insignificant, effects of both EVMS and trip recorders on revenue, again on the order of a 50% advantage for adopters versus non-adopters. One caveat is worth mentioning regarding the results in Table 17. The first-difference estimation scheme is not robust to trends in revenue across sectors. Any trend that begins in the period prior to the introduction of OBCs and continues through the OBC adoption stage can weaken the causal interpretation of the estimates. In particular, this includes the case in which the sectors growing the fastest are the ones investing in EVMS. Unfortunately, given that the data is available for only two points in time, it is not possible to control for trends in the estimation scheme.

Viewed jointly, Tables 16 and 17 consistently display large, positive effects of EVMS on revenue, and in some cases a substantial impact of trip recorder use on revenue. That incentive capabilities can enhance revenue is not unexpected, but the implied magnitude of some of the point estimates is surprising. In theory, robustness tests for this phenomenon can be derived. For example, the incentive capabilities related to non-driving tasks should be greatest for those sectors where loading and unloading and other cargo-handling responsibilities constitute a large fraction of the driver's job. This is true of short-haul drivers, those in the LTL sector, and drivers operating refrigerated trucks.[29] As an empirical matter, though, there is not sufficient power to spot divergent effects of OBC adoption across

[29] Baker and Hubbard (2001) discuss how OBC adoption and the extent of cargo-handling activities interact to determine whether shippers use their own fleets or employ for-hire carriers.

sectors, particularly given the difficulties in identifying the main effects of trip recorder use. Taking the revenue results as a whole, the robustness of the impact of EVMS adoption across different specifications points to an important role for the communication and information enabling qualities in improving resource allocation decisions.

# 5 Conclusion

The creation of new technologies can have drastic effects on how work is organized and performed within firms. Economic theory points to a number of ways in which better monitoring devices can improve worker incentives to align employee actions with employer preferences, as well as enable a more efficient use of resources through a less costly flow of information between economic agents. OBC technology in the trucking industry encompasses both of these benefits to firms. A previous approach in the literature attempts to separately identify the incentive and coordination components through the variation in OBC adoption rates across sectors of the trucking industry. The empirical work presented here illustrates the fragility of this methodology. In particular, it is shown that truck age is a key factor influencing adoption, and that the use of the between model year variation invalidates the previous interpretation of the results. Once the adjustment is made to limit the variation to within model years, the incentive effect is robust, but the coordination effect becomes more difficult to detect. This lack of a finding does not imply that the technology does not proffer coordination improvements, but rather is a strong indication that this methodology is not well-suited to measure the coordination effect.

This paper then improves on previous work by considering the direct impact of the technology on measures associated with incentive and coordination improvements. The results indicate that incentive improvements are manifested in a reduction in vehicle operating

costs, most significantly on expenses related to paying outside mechanics. There is also tentative evidence that the incentive effect can enhance revenue, perhaps due to a reduction in shirking by drivers who have substantial non-driving responsibilities. The revenue impact of EVMS, which has the additional coordination enhancing attribute, is consistent across OLS and fixed effect specifications, though not statistically significant in the fixed effect regressions, and is robust to weighting. The estimates imply that carriers that have outfitted their entire fleet of trucks with EVMS have boosted their revenues on the order of 40% to 50%.

The estimates in this paper point to an enormous benefit of the technology to the trucking industry. Consider a very rough estimate of the total impact of EVMS on revenue. According to the American Trucking Association, the U.S. trucking industry generated about $500 billion of revenue in 1999. Using a point estimate of .45, an eyeball average of the estimates in Tables 16 and 17, this implies a gain of $51 billion per year to the industry due to the diffusion of the technology.[30] As a comparison, Hubbard (2001) estimates that EVMS confer $16 billion per year in benefits in terms of improved capacity utilization of trucks, which is one component of the gross benefit that is estimated using revenue. The revenue measure will capture additional improvements due to more comprehensive customer service via shipment tracking, as well as potential incentive benefits. How accurately the magnitudes of these effects generalize to other industries and technologies is difficult to assess, but the results do suggest that a more fluid transmission of information can help firms more efficiently allocate inputs to their productive uses.

[30]Let $X$ be the counterfactual representing the size of the U.S. trucking industry in 1999 if on-board computers were not available. Assuming that trucks with EVMS generate revenues 45% greater than trucks without EVMS and that 25% of trucks have EVMS installed, then $X$ is defined by the equation: $X(1.45)(.25) + X(1)(.75) = \$500b$. This yields $X = \$449b$, so the gain from the implementation of EVMS is $\$51b$ per year.

## References


[1] Alchian, Armen A., and Harold Demsetz, (1972), 'Production, Information Costs, and Economic Organization,' *American Economic Review*, 62 (5), pp. 777-795.

[2] Baker, George P., and Thomas N. Hubbard, (2000), 'Contractibility and Asset Ownership: On-Board Computers and Governance in U.S. Trucking,' *NBER Working Paper 7634*.

[3] Baker, George P., and Thomas N. Hubbard, (2001), 'Make Versus Buy in Trucking: Asset Ownership, Job Design and Information,' mimeo, University of Chicago Graduate School of Business.

[4] Belman, Dale A., and Kristen A. Monaco, (2001), 'The Effects of Deregulation, Deunionization, Technology, and Human Capital on the Work and Work Lives of Truck Drivers,' *Industrial and Labor Relations Review*, 54 (2A), pp. 502-524.

[5] Bigras, Yvon, Teodor Gabriel Crainic, and Jacques Roy, (1997), 'The Use of Information Technologies in the Motor Carrier Industry,' *Centre de Recherche en Gestion Document 06-97*.

[6] Brynjolfsson, Erik, and Lorin M. Hitt, (2000) 'Beyond Computation: Information Technology, Organizational Transformation, and Business Performance,' *Journal of Economic Perspectives*, 14 (4), pp. 23-48.

[7] Cacciola, Stephen E., (2002), 'Empirical Tests of a Principal-Agent Model: Exploiting On-Board Computer Adoption in the Trucking Industry,' Chapter 2, Ph.D. Dissertation, Yale University.

[8] Chakraborty, Atreya, and Mark Kazarosian, (1999). 'Product Differentiation and the Use of Information Technology: Evidence from the Trucking Industry.' *NBER Working Paper 7222.*

[9] Deaton, Angus, (1985). 'Panel Data from a Time Series of Cross-Sections.' *Journal of Econometrics*, 30, pp. 109-126.

[10] Hayek, F.A., (1945). 'The Use of Knowledge in Society.' *American Economic Review*, 35 (4), pp. 519-530.

[11] Holmstrom, Bengt, (1979). 'Moral Hazard and Observability.' *Bell Journal of Economics*, 10 (1), pp. 74-91.

[12] Holmstrom, Bengt, (1982). 'Moral Hazard in Teams.' *Bell Journal of Economics*, 13 (2), pp. 324-340.

[13] Holmstrom, Bengt, and Paul Milgrom, (1994). 'The Firm as an Incentive System.' *American Economic Review*, 84 (4), pp. 972-991.

[14] Hubbard, Thomas N., (2000). 'The Demand for Monitoring Technologies: The Case of Trucking.' *Quarterly Journal of Economics*, 115 (2), pp. 533-560.

[15] Hubbard Thomas N., (2001). 'Information Decisions and Productivity: On-Board Computers and Capacity Utilization in Trucking,' mimeo, University of Chicago Graduate School of Business.

[16] Lazear, Edward P., (1995), *Personnel Economics*, MIT Press.

[17] Lazear, Edward P., (1999), 'Personnel Economics: Past Lessons and Future Directions,' *Journal of Labor Economics*, 17 (2), pp. 199-236.

[18] Lazear, Edward P., (2000), 'The Future of Personnel Economics,' *The Economic Journal*, 110 (November), pp. F611-F639.

[19] Ouellet, Lawrence J., (1994), *Pedal to the Metal: The Work Lives of Truckers*, Temple University Press.

[20] Phipps, Jeannie L., (2001), 'High- and Low-Tech Techniques for Controlling Fuel Costs,' *Bankrate.com*, June 18, 2001.

[21] Prendergast, Canice, (1996), 'What Happens Within Firms? A Survey of Empirical Evidence on Compensation Policies,' *NBER Working Paper 5802*.

[22] Prendergast, Canice, (1999), 'The Provision of Incentives in Firms,' *Journal of Economic Literature*, 37 (1), pp. 7-63.

[23] Schrodt, Anita, (1989), 'Schneider National Keeps Tabs on Fleet Through Satellites,' *Journal of Commerce*, May 17, 1989, p. 2B.

**Figure 1: OBC Use Rates by Model Year**
**1992 Survey**

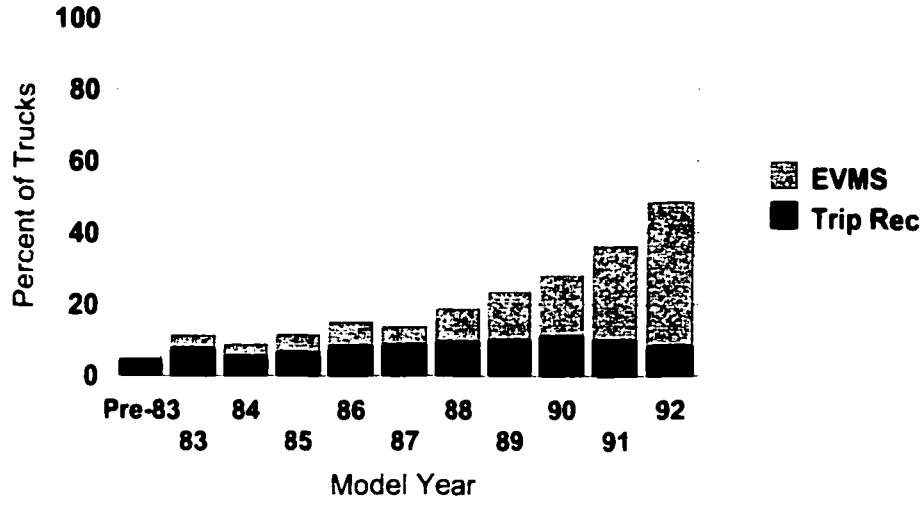


**Figure 2: OBC Use Rates by Model Year**
**1997 Survey**

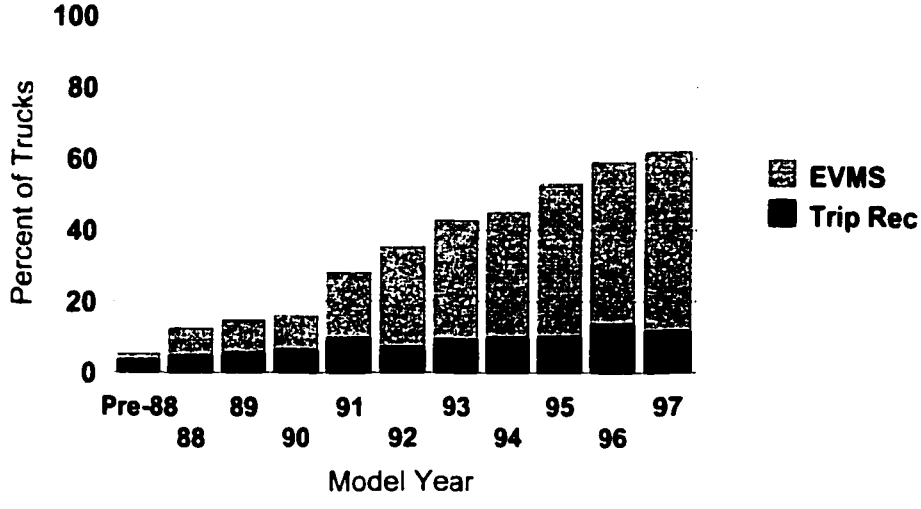

**Table 1**
**Trip Recorder and EVMS Use Rates by Length of Haul**
**(Percent of Sample in Parentheses)**

| | 1987 | | 1992 | | 1997 | |
|---|---|---|---|---|---|---|
| | Trip recorder | EVMS | Trip recorder | EVMS | Trip recorder | EVMS |
| Off-Road | 0 | 0 | 1.80 | 1.25 | 2.83 | 7.89 |
| | (2.52) | | (2.93) | | (2.22) | |
| Local (50 miles or less) | 0 | 0 | 4.20 | 2.46 | 3.89 | 9.29 |
| | (26.53) | | (23.62) | | (21.35) | |
| Short Range (51 to 100 miles) | 0 | 0 | 6.92 | 4.35 | 6.67 | 13.58 |
| | (32.40) | | (17.47) | | (17.63) | |
| Short (101 to 200 miles) | | | 11.28 | 7.67 | 8.50 | 22.04 |
| | | | (13.78) | | (15.17) | |
| Long Range-Medium (201 to 500 miles) | 0 | 0 | 10.82 | 13.54 | 7.15 | 32.58 |
| | (37.70) | | (17.83) | | (18.13) | |
| Long Range (501 miles or more) | | | 6.96 | 21.98 | 12.37 | 39.31 |
| | | | (24.38) | | (25.51) | |
| All Categories | C | 0 | 7.49 | 10.49 | 7.95 | 24.59 |
| (Number of Obs) | (24,778) | | (39,850) | | (25,533) | |

Note: In the 1987 survey, the area of operation categories are more general. The survey options consist of (1) off-road, (2) 50 miles or less, (3) 51-200 miles, and (4) 200 miles or more. Expansion factors provided by the Census are used as weights.

**Table 2**
**Trip Recorder and EVMS Use Rates by Trailer Type**
**(Percent of Sample in Parentheses)**

| | 1987 | | 1992 | | 1997 | |
|---|---|---|---|---|---|---|
| | Trip recorder | EVMS | Trip recorder | EVMS | Trip recorder | EVMS |
| Tank Truck | 0 (8.46) | 0 | 14.23 (8.93) | 7.49 | 8.78 (7.86) | 27.98 |
| Refrigerated Van | 0 (9.66) | 0 | 13.46 (11.40) | 20.88 | 16.85 (10.81) | 34.60 |
| Dry Cargo Van | 0 (31.72) | 0 | 7.78 (32.70) | 14.27 | 9.03 (36.46) | 30.90 |
| Platform | 0 (26.98) | 0 | 4.46 (22.43) | 5.57 | 4.73 (21.60) | 15.69 |
| Dump Truck | 0 (6.84) | 0 | 3.75 (7.53) | 3.59 | 4.00 (7.17) | 12.15 |
| Grain Bodies | 0 (3.35) | 0 | 2.06 (3.88) | 3.35 | 1.58 (4.54) | 8.87 |
| Other | 0 (13.00) | 0 | 5.48 (13.12) | 6.68 | 4.56 (11.55) | 16.18 |
| All Categories (Number of Obs) | 0 (24,989) | 0 | 7.49 (39,850) | 10.49 | 7.95 (25,533) | 24.59 |

Note: Expansion factors provided by the Census are used as weights.

**Table 3**
**Trip Recorder and EVMS Use Rates by Type of Service**
**(Percent of Sample in Parentheses)**

| | 1987 | | 1992 | | 1997 | |
|---|---|---|---|---|---|---|
| | Trip recorder | EVMS | Trip recorder | EVMS | Trip recorder | EVMS |
| Truckload | | | 6.82 (36.74) | 17.18 | 10.98 (35.95) | 33.39 |
| Less-than-truckload | | | 5.16 (10.15) | 10.14 | 4.59 (11.07) | 23.52 |
| Private Fleet | | | 8.83 (53.11) | 5.84 | 7.24 (52.99) | 17.29 |
| All Categories (Number of Obs) | | | 7.64 (37,559) | 10.72 | 8.28 (23,914) | 24.14 |

Notes: This information is not available in the 1987 survey. The truckload and less-than-truckload distinction is only appropriate for trucks employed by for-hire firms. Expansion factors provided by the Census are used as weights.

**Table 4**
**Linear Probability Model Estimates of OBC Use, and Trip Recorder and EVMS Use Conditional on OBC Use, 1992 Survey (*Pooled Model Years*)**

| | **OBC Use = 1 Non-OBC Use = 0** | | **Conditional on OBC Use** (EVMS = 1, Trip Rec. = 0) | |
|---|---|---|---|---|
| | Estimate | Stand. Error | Estimate | Stand. Error |
| 50-100 Miles | **.031** | **.007** | .018 | .039 |
| 100-200 Miles | **.086** | **.009** | .016 | .038 |
| 200-500 Miles | **.131** | **.009** | **.107** | **.036** |
| Over 500 Miles | **.152** | **.010** | **.157** | **.038** |
| (Omitted is <50 Miles) | | | | |
| Tank Truck | .028 | .015 | **-.177** | **.044** |
| Refrigerated Van | **.099** | **.013** | .051 | .028 |
| Platform Trailer | **-.024** | **.009** | -.043 | .034 |
| Specialized Trailer | -.010 | .009 | .018 | .034 |
| (Omitted is Dry Van) | | | | |
| Truckload | **.137** | **.012** | .070 | .037 |
| (Omitted is LTL) | | | | |
| Private Fleet | **.105** | **.012** | **-.203** | **.038** |
| Contract Carriage | **-.019** | **.009** | **-.062** | **.020** |
| (Omitted is Common) | | | | |
| Intrastate Operation | **-.038** | **.010** | **-.104** | **.041** |
| Owner-Operator | **-.110** | **.009** | -.040 | .042 |
| Private Refuel Facility | **.067** | **.006** | **-.096** | **.020** |
| Exempt Carrier | -.001 | .014 | **-.142** | **.058** |
| Fleet Size 25-99 | **.041** | **.007** | -.053 | .029 |

| | | | | |
|---|---|---|---|---|
| Fleet Size 100-499 | **.139** | **.010** | **-.093** | **.029** |
| Fleet Size 500-999 | **.193** | **.017** | **-.070** | **.034** |
| Fleet Size 1000-4999 | **.225** | **.014** | -.039 | .031 |
| Fleet Size 5000-10000 | **.274** | **.022** | **.106** | **.048** |
| Fleet Size over 10000 | **.170** | **.016** | **.117** | **.046** |
| Number of Obs | 35,253 | --- | 6,023 | 3,311 EVMS<br>2,712 TR |

Notes: Estimates in bold indicate significance at the 5% level. Empirical propositions are highlighted. Expansion factors provided by the Census are used as weights. Other covariates include principal product hauled and base state.

**Table 5**
**Linear Probability Model Estimates of OBC Use**
**Coefficients on the Truckload Variable by Age of Vehicle, 1992 Sample**

| Age of Vehicle: | Trip Recorder = 1 (Non-Trip Recorder = 0) | EVMS = 1 (Non-EVMS = 0) |
|---|---|---|
| New (Model Year 1992) | -.003 (.020) | **.195 (.040)** |
| 1 (Model Year 1991) | .009 (.031) | **.098 (.044)** |
| 2 (Model Year 1990) | .054 (.029) | **.181 (.031)** |
| 3 (Model Year 1989) | -.024 (.030) | **.126 (.024)** |
| 4 (Model Year 1988) | .031 (.024) | **.119 (.024)** |
| 5 (Model Year 1987) | .040 (.022) | **.038 (.018)** |
| 6 (Model Year 1986) | -.009 (.028) | **.062 (.029)** |
| 7 (Model Year 1985) | **.072 (.031)** | .027 (.018) |
| 8 (Model Year 1984) | **.065 (.024)** | -.007 (.018) |
| 9 (Model Year 1983) | **.139 (.067)** | -.045 (.064) |
| 10 or Older (Model Year Pre-1982) | .012 (.017) | **.021 (.007)** |

Notes: Standard errors are in parentheses. Estimates in bold indicate significance at the 5% level. Expansion factors provided by the Census are used as weights. Other covariates include length of haul, trailer type, private fleet, contract carriage, intrastate operation, owner-operator, private refuel facility, exempt carrier, fleet size, principal product carried, and base state.

**Table 6**
**Truck Age Regressions as a Function of Truck Characteristics, By Survey Year**

| | 1987 Survey | 1992 Survey | 1997 Survey |
|---|---|---|---|
| 50-100 Miles | -.896 | -.869 | -.731 |
| | (.070) | (.073) | (.094) |
| 100-200 Miles | --- | -1.309 | -1.463 |
| | | (.079) | (.101) |
| 200-500 Miles | -2.195 | -1.951 | -2.178 |
| | (.082) | (.077) | (.103) |
| Over 500 Miles | --- | -3.126 | -3.104 |
| | | (.080) | (.103) |
| Tank Truck | -.151 | -.419 | -.166 |
| | (.150) | (.120) | (.161) |
| Refrigerated Van | -.345 | -.327 | -.292 |
| | (.116) | (.093) | (.122) |
| Platform Trailer | .998 | .773 | .637 |
| | (.088) | (.079) | (.103) |
| Specialized Trailer | .680 | .374 | .353 |
| | (.095) | (.080) | (.109) |
| Truckload | --- | -.192 | -.637 |
| | | (.099) | (.112) |
| Private Fleet | .567 | .010 | -.117 |
| | (.083) | (.103) | (.115) |
| Contract Carriage | .039 | -.213 | -.471 |
| | (.092) | (.069) | (.081) |
| Intrastate Operation | .936 | .450 | .864 |
| | (.090) | (.086) | (.117) |
| Owner-Operator | .805 | .995 | .941 |
| | (.089) | (.084) | (.106) |
| Private Refuel Facility | --- | .006 | -.053 |
| | | (.051) | (.067) |
| Exempt Carrier | -.023 | -.272 | -.614 |
| | (.149) | (.145) | (.234) |
| Fleet Size 25-99 | -.709 | -.969 | -1.011 |
| | (.069) | (.066) | (.091) |
| Fleet Size 100-499 | -1.242 | -1.621 | -1.823 |
| | (.085) | (.076) | (.092) |

| | | | |
|---|---|---|---|
| Fleet Size 500-999 | -1.477 | -1.808 | -2.128 |
| | (.113) | (.113) | (.137) |
| Fleet Size 1000-4999 | --- | -1.855 | -1.978 |
| | | (.103) | (.123) |
| Fleet Size 5000-10000 | --- | -1.709 | -2.012 |
| | | (.168) | (.183) |
| Fleet Size over 10000 | --- | -2.416 | -2.455 |
| | | (.112) | (.121) |
| Number of Obs | 24,989 | 35,026 | 22,122 |

Notes: Standard errors are in parentheses. Expansion factors provided by the Census are used as weights. Other covariates include principal product hauled and base state. Dashes in the place of estimates indicate that certain categories and variables are not available in the 1987 survey.

**Table 7**
**Linear Probability Model Estimates of OBC Use, and Trip Recorder and EVMS Use Conditional on OBC Use, 1992 Survey (*New Trucks - Model Year 1992*)**

| | **OBC Use = 1 Non-OBC Use = 0** | | **Conditional on OBC Use (EVMS = 1, Trip Rec. = 0)** | |
|---|---|---|---|---|
| | Estimate | Stand. Error | Estimate | Stand. Error |
| 50-100 Miles | .092 | .054 | .063 | .091 |
| 100-200 Miles | **.167** | **.052** | -.122 | .099 |
| 200-500 Miles | **.182** | **.051** | .028 | .089 |
| Over 500 Miles | **.188** | **.053** | .021 | .098 |
| (Omitted is <50 Miles) | | | | |
| Tank Truck | .066 | .065 | -.072 | .073 |
| Refrigerated Van | .028 | .040 | -.025 | .043 |
| Platform Trailer | -.081 | .050 | -.012 | .052 |
| Specialized Trailer | **-.118** | **.047** | -.080 | .067 |
| (Omitted is Dry Van) | | | | |
| Truckload | **.190** | **.041** | .005 | .054 |
| (Omitted is LTL) | | | | |
| Private Fleet | -.002 | .046 | **-.206** | **.065** |
| Contract Carriage | -.035 | .030 | **-.075** | **.027** |
| (Omitted is Common) | | | | |
| Intrastate Operation | **-.186** | **.055** | -.009 | .056 |
| Owner-Operator | **-.164** | **.058** | -.134 | .089 |
| Number of Obs | 3,097 | --- | 1,472 | 1,151 EVMS<br>321 TR |

Notes: Estimates in bold indicate significance at the 5% level. Empirical propositions are highlighted. Expansion factors provided by the Census are used as weights. Other covariates include private refuel facility, exempt carrier, fleet size, principal product hauled, and base state.

**Table 8**
**Linear Probability Model Estimates of OBC Use, and Trip Recorder and EVMS Use Conditional on OBC Use, 1992 Survey (*Trucks Ages 1 to 4*)**

| | **OBC Use = 1 Non-OBC Use = 0** | | **Conditional on OBC Use (EVMS = 1, Trip Rec. = 0)** | |
|---|---|---|---|---|
| | Estimate | Stand. Error | Estimate | Stand. Error |
| 50-100 Miles | .027 | .018 | -.096 | .059 |
| 100-200 Miles | **.101** | **.020** | -.078 | .062 |
| 200-500 Miles | **.142** | **.019** | -.004 | .058 |
| Over 500 Miles | **.164** | **.019** | .062 | .059 |
| (Omitted is <50 Miles) | | | | |
| Tank Truck | -.021 | .025 | **-.227** | **.058** |
| Refrigerated Van | **.108** | **.021** | .044 | .040 |
| Platform Trailer | -.029 | .021 | .031 | .048 |
| Specialized Trailer | **-.041** | **.019** | .025 | .053 |
| (Omitted is Dry Van) | | | | |
| Truckload | **.156** | **.020** | .084 | .051 |
| (Omitted is LTL) | | | | |
| Private Fleet | **.140** | **.021** | **-.139** | **.055** |
| Contract Carriage | **-.064** | **.015** | -.039 | .028 |
| (Omitted is Common) | | | | |
| Intrastate Operation | -.010 | .021 | -.029 | .055 |
| Owner-Operator | **-.086** | **.019** | .048 | .045 |
| Number of Obs | 11,018 | --- | 2,876 | 1,651 EVMS<br>1,225 TR |

Notes: Estimates in bold indicate significance at the 5% level. Empirical propositions are highlighted. Expansion factors provided by the Census are used as weights. Other covariates include private refuel facility, exempt carrier, fleet size, principal product hauled, and base state.

**Table 9**
**Linear Probability Model Estimates of OBC Use, and Trip Recorder and EVMS Use Conditional on OBC Use, 1992 Survey (*Trucks Older than 5 Years*)**

| | **OBC Use = 1 Non-OBC Use = 0** | | **Conditional on OBC Use (EVMS = 1, Trip Rec. = 0)** | |
|---|---|---|---|---|
| | Estimate | Stand. Error | Estimate | Stand. Error |
| 50-100 Miles | **.023** | **.007** | -.022 | .047 |
| 100-200 Miles | **.065** | **.010** | -.043 | .046 |
| 200-500 Miles | **.103** | **.011** | -.024 | .046 |
| Over 500 Miles (Omitted is <50 Miles) | **.033** | **.011** | -.029 | .065 |
| Tank Truck | **.062** | **.018** | -.111 | .093 |
| Refrigerated Van | **.087** | **.017** | .010 | .051 |
| Platform Trailer | -.002 | .009 | -.128 | .055 |
| Specialized Trailer (Omitted is Dry Van) | .017 | .010 | -.035 | .060 |
| Truckload (Omitted is LTL) | **.069** | **.013** | -.069 | .071 |
| Private Fleet | **.076** | **.015** | **-.305** | **.071** |
| Contract Carriage (Omitted is Common) | **.021** | **.010** | -.042 | .047 |
| Intrastate Operation | **-.028** | **.010** | **-.161** | **.057** |
| Owner-Operator | **-.051** | **.008** | -.031 | .091 |
| Number of Obs | 21,138 | --- | 1,675 | 509 EVMS<br>1166 TR |

Notes: Estimates in bold indicate significance at the 5% level. Empirical propositions are highlighted. Expansion factors provided by the Census are used as weights. Other covariates include private refuel facility, exempt carrier, fleet size, principal product hauled, and base state.

**Table 10**
**Cross Section OLS Regressions of Log Fuel Costs per Ton-Mile on On-Board Computer Use**

| Weighting Scheme: | Unweighted | Unweighted | Weighted | Weighted |
|---|---|---|---|---|
| EVMS | -.820<br>(.555) | -.545<br>(.596) | .030<br>(.458) | -.130<br>(.533) |
| Trip Recorder | .177<br>(1.046) | -.218<br>(1.088) | .289<br>(.584) | -.236<br>(.733) |
| Main Effects for Cells Included? | No | Yes | No | Yes |
| Test for Joint Significance of Main Effects for Cells: (p-value) | ----- | F(55, 209) = 2.10<br>(0.0001) | ----- | F(55, 209) = 2.62<br>(0.0000) |
| Additional Controls | Truck Age<br>Fleet Size | Truck Age<br>Fleet Size | Truck Age<br>Fleet Size | Truck Age<br>Fleet Size |
| Number of Cells | 278 | 278 | 278 | 278 |

Notes: Each column represents a separate regression. Standard errors are in parentheses. Cells in bold indicate significance at the 10% level. Cells are constructed by base state, cargo type, and length of haul. The inclusion of main effects for the cells consists of dummy variables for base state, cargo type, and length of haul added to the regression. The weighting scheme used is (n99+n97), where n99 is the number of firms in a given cell in the financial data in 1999, and n97 is the number of trucks in a given cell in the on-board computer use data in 1997.

**Table 11**
**Regressions of the Change in Log Fuel Costs per Ton-Mile on On-Board Computer Use**

| Weighting Scheme: | Unweighted | Unweighted | Weighted | Weighted |
|---|---|---|---|---|
| EVMS | -.392<br>(.650) | -.887<br>(.804) | .542<br>(.567) | .142<br>(.728) |
| Trip Recorder | .435<br>(1.153) | -.440<br>(1.296) | .216<br>(.696) | -1.01<br>(.939) |
| Main Effects for Cells Included? | No | Yes | No | Yes |
| Test for Joint Significance of Main Effects for Cells: (p-value) | ----- | F(55, 171) = 1.35<br>(0.0740) | ----- | F(55, 171) = 1.62<br>(0.0101) |
| Additional Controls | Truck Age<br>ΔFleet Size | Truck Age<br>ΔFleet Size | Truck Age<br>ΔFleet Size | Truck Age<br>ΔFleet Size |
| Number of Cells | 278 | 278 | 278 | 278 |

Notes: Each column represents a separate regression. Standard errors are in parentheses. Cells in bold indicate significance at the 10% level. Cells are constructed by base state, cargo type, and length of haul. The inclusion of main effects for the cells consists of dummy variables for base state, cargo type, and length of haul added to the regression. The weighting scheme used is (n89+n99+n97), where n89 is the number of firms in a given cell in the financial data in 1989, n99 is the number of firms in a given cell in the financial data in 1999, and n97 is the number of trucks in a given cell in the on-board computer use data in 1997.

**Table 12**
**Cross Section OLS Regressions of Log Outside Maintenance Costs per Ton-Mile on On-Board Computer Use**

| Weighting Scheme: | Unweighted | Unweighted | Weighted | Weighted |
|---|---|---|---|---|
| EVMS | **-1.203** | -.859 | -.422 | -.042 |
| | **(.559)** | (.625) | (.502) | (.613) |
| Trip Recorder | **-2.223** | **-1.971** | -.790 | -.550 |
| | **(1.081)** | **(1.144)** | (.638) | (.834) |
| Main Effects for Cells Included? | No | Yes | No | Yes |
| Test for Joint Significance of Main Effects for Cells: (p-value) | ----- | F(54, 202) = 1.80 (0.0019) | ----- | F(54, 202) = 2.16 (0.0001) |
| Additional Controls | Truck Age Fleet Size | Truck Age Fleet Size | Truck Age Fleet Size | Truck Age Fleet Size |
| Number of Cells | 270 | 270 | 270 | 270 |

Notes: Each column represents a separate regression. Standard errors are in parentheses. Cells in bold indicate significance at the 10% level. Cells are constructed by base state, cargo type, and length of haul. The inclusion of main effects for the cells consists of dummy variables for base state, cargo type, and length of haul added to the regression. The weighting scheme used is (n99+n97), where n99 is the number of firms in a given cell in the financial data in 1999, and n97 is the number of trucks in a given cell in the on-board computer use data in 1997.

**Table 13**
**Regressions of the Change in Log Total Outside Maintenance Costs per Ton-Mile on On-Board Computer Use**

| Weighting Scheme: | Unweighted | Unweighted | Weighted | Weighted |
|---|---|---|---|---|
| EVMS | **-1.878** | -1.166 | .319 | .354 |
| | **(.759)** | (1.031) | (.703) | (1.009) |
| Trip Recorder | **-3.226** | -2.383 | -.909 | -.854 |
| | **(1.294)** | (1.523) | (.795) | (1.163) |
| Main Effects for Cells Included? | No | Yes | No | Yes |
| Test for Joint Significance of Main Effects for Cells: (p-value) | ----- | F(53, 137) = 1.14 (0.2741) | ----- | F(53, 137) = 1.49 (0.0342) |
| Additional Controls | Truck Age ΔFleet Size | Truck Age ΔFleet Size | Truck Age ΔFleet Size | Truck Age ΔFleet Size |
| Number of Cells | 270 | 270 | 270 | 270 |

Notes: Each column represents a separate regression. Standard errors are in parentheses. Cells in bold indicate significance at the 10% level. Cells are constructed by base state, cargo type, and length of haul. The inclusion of main effects for the cells consists of dummy variables for base state, cargo type, and length of haul added to the regression. The weighting scheme used is (n89+n99+n97), where n89 is the number of firms in a given cell in the financial data in 1989, n99 is the number of firms in a given cell in the financial data in 1999, and n97 is the number of trucks in a given cell in the on-board computer use data in 1997.

**Table 14**
**Cross Section OLS Regressions of Log Total Operating Supply Costs per Ton-Mile on On-Board Computer Use**

| Weighting Scheme: | Unweighted | Unweighted | Weighted | Weighted |
|---|---|---|---|---|
| EVMS | -.683<br>(.511) | -.365<br>(.551) | -.234<br>(.406) | -.256<br>(.468) |
| Trip Recorder | -.281<br>(.958) | -.370<br>(.995) | -.216<br>(.517) | .246<br>(.642) |
| Main Effects for Cells Included? | No | Yes | No | Yes |
| Test for Joint Significance of Main Effects for Cells: (p-value) | ----- | F(55, 213) = 2.08<br>(0.0001) | ----- | F(55, 213) = 2.72<br>(0.0000) |
| Additional Controls | Truck Age<br>Fleet Size | Truck Age<br>Fleet Size | Truck Age<br>Fleet Size | Truck Age<br>Fleet Size |
| Number of Cells | 282 | 282 | 282 | 282 |

Notes: Each column represents a separate regression. Standard errors are in parentheses. Cells in bold indicate significance at the 10% level. Cells are constructed by base state, cargo type, and length of haul. The inclusion of main effects for the cells consists of dummy variables for base state, cargo type, and length of haul added to the regression. The weighting scheme used is (n99+n97), where n99 is the number of firms in a given cell in the financial data in 1999, and n97 is the number of trucks in a given cell in the on-board computer use data in 1997.

**Table 15**
**Regressions of the Change in Log Total Operating Supply Costs per Ton-Mile on On-Board Computer Use**

| Weighting Scheme: | Unweighted | Unweighted | Weighted | Weighted |
|---|---|---|---|---|
| EVMS | -.480<br>(.615) | -.990<br>(.784) | .286<br>(.489) | -.122<br>(.642) |
| Trip Recorder | .076<br>(1.083) | -.560<br>(1.244) | .113<br>(.598) | -.359<br>(.824) |
| Main Effects for Cells Included? | No | Yes | No | Yes |
| Test for Joint Significance of Main Effects for Cells: (p-value) | ----- | F(55, 178) = 1.11<br>(0.3067) | ----- | F(55, 178) = 1.40<br>(0.0536) |
| Additional Controls | Truck Age<br>ΔFleet Size | Truck Age<br>ΔFleet Size | Truck Age<br>ΔFleet Size | Truck Age<br>ΔFleet Size |
| Number of Cells | 282 | 282 | 282 | 282 |

Notes: Each column represents a separate regression. Standard errors are in parentheses. Cells in bold indicate significance at the 10% level. Cells are constructed by base state, cargo type, and length of haul. The inclusion of main effects for the cells consists of dummy variables for base state, cargo type, and length of haul added to the regression. The weighting scheme used is (n89+n99+n97), where n89 is the number of firms in a given cell in the financial data in 1989, n99 is the number of firms in a given cell in the financial data in 1999, and n97 is the number of trucks in a given cell in the on-board computer use data in 1997.

**Table 16**
**Cross Section OLS Regressions of Log Revenue on On-Board Computer Use**

| Weighting Scheme: | Unweighted | Unweighted | Weighted | Weighted |
|---|---|---|---|---|
| EVMS | **.302**<br>**(.150)** | **.498**<br>**(.161)** | .132<br>(.126) | **.285**<br>**(.149)** |
| Trip Recorder | .094<br>(.265) | **.476**<br>**(.277)** | -.015<br>(.159) | -.106<br>(.203) |
| Main Effects for Cells Included? | No | Yes | No | Yes |
| Test for Joint Significance of Main Effects for Cells: (p-value) | ----- | F(57, 250) = 1.94<br>(0.0003) | ----- | F(57, 250) = 2.34<br>(0.0000) |
| Additional Controls | Truck Age<br>Fleet Size | Truck Age<br>Fleet Size | Truck Age<br>Fleet Size | Truck Age<br>Fleet Size |
| Number of Cells | 321 | 321 | 321 | 321 |

Notes: Each column represents a separate regression. Standard errors are in parentheses. Cells in bold indicate significance at the 10% level. Cells are constructed by base state, cargo type, and length of haul. The inclusion of main effects for the cells consists of dummy variables for base state, cargo type, and length of haul added to the regression. The weighting scheme used is (n99+n97), where n99 is the number of firms in a given cell in the financial data in 1999, and n97 is the number of trucks in a given cell in the on-board computer use data in 1997.

**Table 17**
**Regressions of the Change in Log Revenue on On-Board Computer Use**

| Weighting Scheme: | Unweighted | Unweighted | Weighted | Weighted |
|---|---|---|---|---|
| EVMS | .095<br>(.262) | .444<br>(.300) | .288<br>(.267) | .494<br>(.315) |
| Trip Recorder | -.158<br>(.449) | .062<br>(.503) | .481<br>(.333) | .569<br>(.425) |
| Main Effects for Cells Included? | No | Yes | No | Yes |
| Test for Joint Significance of Main Effects for Cells: (p-value) | ----- | F(56, 222) = 1.37<br>(0.0567) | ----- | F(56, 222) = 2.15<br>(0.0000) |
| Additional Controls | Truck Age<br>ΔFleet Size | Truck Age<br>ΔFleet Size | Truck Age<br>ΔFleet Size | Truck Age<br>ΔFleet Size |
| Number of Cells | 321 | 321 | 321 | 321 |

Notes: Each column represents a separate regression. Standard errors are in parentheses. Cells in bold indicate significance at the 10% level. Cells are constructed by base state, cargo type, and length of haul. The inclusion of main effects for the cells consists of dummy variables for base state, cargo type, and length of haul added to the regression. The weighting scheme used is (n89+n99+n97), where n89 is the number of firms in a given cell in the financial data in 1989, n99 is the number of firms in a given cell in the financial data in 1999, and n97 is the number of trucks in a given cell in the on-board computer use data in 1997.

# Chapter 2
# Empirical Tests of a Principal-Agent Model: Exploiting On-Board Computer Adoption in the Trucking Industry

## 1 Introduction

Agency theory is a workhorse model in microeconomics, implemented in numerous situations to describe the interactions of economic parties who are involved in long-term relationships. In the employment context, the field of personnel economics has borrowed liberally from this previous theoretical work, and has provided substantial insights into many aspects of the behavior of individuals within firms. A particular emphasis has been placed on the design of compensation policies and the optimal structure of incentives.[1] While this theoretical work comprises a rich class of descriptive models, the testing and evaluation of these models has, until recently, been largely inadequate. To help address this deficiency, this paper undertakes an empirical study of a key issue in personnel economics, how the ability of employers to directly monitor the magnitudes and directions of employee effort changes the nature of the employment relationship, and, ultimately, employee actions. The introduction of a sophisticated monitoring technology in the trucking industry provides a unique opportunity to devise tests of some of the central tenets of the theory, namely how risk (in the form of imprecise performance measurement) affects the structure of employee contracts, and how employees respond to changes in the monitoring and incentive environment.

The theoretical foundations of employer and employee interactions have a long history in

[1] For comprehensive discussions of the range of issues addressed by personnel economics, as well as the field's goals and future directions, see Lazear (1995, 1999a, 2000a).

the agency literature.[2] Of particular relevance to this paper, the early work of Alchian and Demsetz (1972) recognizes the value of monitoring in preventing employees from shirking in production. Holmstrom (1979, 1982) formalizes the idea that the information provided by monitoring can be used to improve contractual agreements in order to resolve, at least partially, the moral hazard problem. Other refinements and extensions to the optimal contracting literature of note include Holmstrom and Milgrom (1987) who specify the assumptions necessary for the optimality of *linear* incentive contracts, and Holmstrom and Milgrom (1991) who consider multiple tasks in the agent's job and the role of job design in creating optimal incentives.

As for the inadequacy of the empirical counterpart to this theoretical literature, Prendergast (1996) points to two main reasons: first, the absence of data on contracts and performance, and second, the difficulty in empirically distinguishing between theoretical models. This paper directly addresses these two issues by exploiting the introduction of on-board computer (OBC) technology in the trucking industry. Starting in the late 1980's, trucking firms have the ability to install small computers, called OBCs, on individual trucks. OBCs collect a wealth of information on how a truck is operated by the driver, including episodes of driving at high rates of speed, over-revving of the engine, and excessive idling. Trucking firms value this information because poor driving behavior stresses the truck and results in additional costs borne by the firm, in the form of more frequent breakdowns, reduced truck life expectancy, and lower fuel efficiency. The data collected by OBCs and subsequently processed by firms can be used to make driver contracts contingent on performance, providing incentives for drivers to operate trucks in a manner more aligned with

[2]Optimal incentive contracts were initially studied in a variety of contexts, including insurance, tax policies, and employment contracts. Early contributions to this literature include Mirrlees (1971), Spence and Zechkhauser (1971), and Ross (1973).

their employers' preferences.

The empirical work in this paper tackles two issues. First, I use variation in the adoption rates of OBCs across sectors of the trucking industry to test a very specific prediction in agency theory. The theoretical model of contract choice presented here is the Holmstrom-Milgrom (1991) multitask principal-agent model adapted to the trucking firm and company driver relationship. In the trucking context, one of the driver's tasks (the on-time arrival task) is easily observed without an OBC, but the effort directed towards his other task (the operating behavior task) becomes much more precisely measured upon adoption of an OBC. The model predicts that as the variance (or 'noise') of observed output falls for one task, then the incentive intensities should increase for *both* tasks. Since OBCs allow for a reduction in this variance, sectors of the industry with high OBC use are predicted to use 'high-powered' contracts that tie compensation more closely to driver performance. Using cell-level data constructed from combining truck-level OBC data from the Census of Transportation with a data set rich in firm-level information on driver contracts, the evidence in a single cross-section is mixed. The incentive intensity associated with the operating behavior task does appear to increase upon adoption, but the incentive intensity for the on-time arrival task does not change with adoption. This latter empirical finding may be due to low power resulting from a small sample size.

Given the more precise measurement of driver effort under adoption of an OBC, and the attendant changes in the observed (and unobserved) components of the compensation contract, in a following section I then test for the impact of monitoring on driver behavior. The theoretical model predicts that monitoring of drivers should lead them to adjust the mean and variance of their driving speed to a level that is more compatible with their employers' objectives. This suggests two avenues to detect changes in driver behavior: first, an OBC should increase the life expectancy of a truck, and second, an OBC should increase a truck's

fuel efficiency. By constructing a synthetic panel (as in Deaton (1985)) using three waves of Census of Transportation data, I am able to carefully consider the timing and endogeneity issues involved in uncovering the causal relationships between the two dependent variables of interest and OBC adoption. The empirical results indicate that adoption increases truck life expectancy by roughly one year, a finding that is consistent across first-difference and instrumental variables specifications. Also, fuel efficiency is improved by nearly 3% upon adoption of an OBC. These results provide direct evidence that monitoring significantly alters driver behavior.

The remainder of the paper is structured as follows. Section 2 discusses some of the difficulties involved in producing meaningful empirical work in personnel economics, and briefly reviews the previous empirical literature. Section 3 provides a description of the trucking industry and OBC technology. In Section 4, I present a theoretical model of contract choice. Section 5 tests a prediction of the theoretical model by estimating the relationship between OBC adoption and contract structure. Section 6 presents the results for truck life expectancy and fuel efficiency, which capture how workers respond to changes in monitoring. Finally, Section 7 concludes.

## 2 Empirical Strategies in Personnel Economics

The lag of empirical analysis behind theoretical development in personnel economics can be explained by several different factors. In a survey of the empirical evidence on compensation policies, Prendergast (1996) outlines a number of common problems. The first issue he addresses is not on the testing side of the research venture per se, but is a problem that he refers to as one of 'theoretical identification'. While incentive theories deliver a number of testable predictions, some of these predictions are not empirically distinguishable from

the implications of other classes of models. Another issue is what Prendergast calls the 'empirical identification' problem, and is simply the standard endogeneity problem familiar to all empirical researchers. In the incentive theory context, the problem is that the variation in compensation contracts across firms, as well as the assignment of contracts to workers within firms, is not done randomly, but rather is a choice. So in order to establish a causal relationship between contract structure and some outcome variable, the empirical methodology must deal with this selection issue. Finally, a reality that researchers in personnel economics must face is the existence of very few appropriate data sets. Information is needed both on contracts offered to workers (i.e. documentation of how measures of performance are tied to pay), as well as relevant measures of employee performance and behavior. Until recently such data has been very limited, hampering the empirical progress in this field.

Conditional on the availability of appropriate data, two main conceptual approaches have been used to test the empirical implications of agency theory. The first is to examine whether observed contracts vary in a way that is predicted by the theory. For example, the standard principal-agent model predicts that a greater sensitivity of pay to performance is negatively associated with the agent's risk aversion (reflecting the trade-off between insurance and incentives in contracts), the variance in the stochastic portion of output, and the marginal cost of agent effort. So if, say, worker output is more precisely measured so that the variance of output decreases, do we see the use of higher-powered incentives? The second method is to study whether incentives matter in the determination of agent behavior. That is, does tying pay more closely to performance increase productivity? As Prendergast (1999) points out, this approach is not a test of the optimal contract solution, but rather it is a necessary ingredient for the theory that agents do in fact respond to incentives.

Several excellent articles exist that survey the empirical work dealing with tests of agency

theory (see Baker, Jensen, and Murphy (1988), Prendergast (1996, 1999), Gibbons (1997), and Chiappori and Salanié (2000)). I will therefore attempt only a brief summary of the relevant literature. The evidence concerning the optimality of observed contracts is somewhat unclear. A common finding of the empirical literature is that pay is not very sensitive to performance. In an early study, Medoff and Abraham (1981) examine the pay of professional and managerial workers in two large manufacturing firms. Although both companies use subjective ratings of their workers to establish bonuses, the authors find little difference in earnings resulting from superior performance, as well as a reluctance by evaluators to give poor ratings. Many studies examine the compensation structures of chief executive officers. A seminal paper in this context is Jensen and Murphy (1990). They estimate that a $1,000 increase in a firm's value results in a $3.25 increase in manager's wealth, and argue that this incentive intensity is inadequate.[3] Baker, Gibbs, and Holmstrom (1994a, 1994b) obtained confidential records of all management personnel employed by a medium-sized U.S. firm in the services industry. Their data on bonus rates indicate that worker performance does not significantly affect total compensation, with the median bonus being less than 10% of salary for non-executive employees. While there are many possible explanations for the prevalent finding of a low intensity of pay incentives, certainly one important factor is that high intensity schemes often have unintended consequences. For example, consider the oft-cited case of a manufacturer who is paid based on the *quantity* of his output and ends up shirking on the *quality* dimension of production. As stated by Gibbons (1997: 10), "When measured performance omits important dimensions of total contribution, firms understand that they will 'get what they pay for,' and so may choose weak incentives in preference to

[3]Both the interpretation and magnitude of this estimate have been challenged in later studies. Haubrich (1994) shows that even fairly low values for an agent's risk aversion are consistent with this estimate. Hall and Liebman (1998), using more recent data than Jensen and Murphy, estimate a greater sensitivity of pay to performance, much of it due to the value of stock options.

strong but frequently dysfunctional incentives." Brickley and Zimmerman (2002) provide one of the first empirical documentations of this substitution of effort by workers across tasks by analyzing changes in incentives for faculty in a business school. Upon an increase in the emphasis placed on teaching relative to research, average teacher ratings improved while research output decreased. The findings on how contracts vary with relevant parameters, such as the agent's risk aversion and the noisiness of performance measures, is mixed. Higgs (1973), Garen (1994), and Gaynor and Gertler (1995) provide tentative evidence that increases in risk drive down the sensitivity of pay to performance. Aggarwal and Samwick (1999) find that CEO's sensitivity of pay to performance does vary with the volatility of stock returns, while Garen (1994) finds little evidence that the noisiness of performance measures affect compensation contracts.

As for whether, and to what extent, incentives matter, a number of recent papers have found significant effects of contractual form on worker behavior. Perhaps the gold-standard data-wise is Lazear (2000b). He analyzes data from Safelite Glass Corp., the largest U.S. installer of automobile glass windows. Following a change in management in 1994, Safelite gradually shifted the payments to its glass installers from an hourly wage to a piece rate schedule. Lazear finds that the introduction of piece rates led to a 36% overall gain in productivity. While about half of the total increase can be attributed to the incentive effect, the Lazear study is also valuable in that it emphasizes the role of the 'selection effect' of contracts, whereby upon the adoption of piece rates, poor quality workers leave the firm and are replaced by higher productivity workers.[4] This selection effect accounts for the other half of the total productivity gain. Shearer (2000) uses data from a field experiment in which workers were randomly assigned to plant trees under either a fixed wage or a piece

[4]Lazear (1999b) develops a theoretical framework that illustrates how a small sensitivity of pay to performance can generate substantial selection and sorting effects.

rate. He finds an incentive effect on the order of 20%, a number similar in magnitude to Lazear (2000b). Fernie and Metcalf (1999) find that jockeys in horse races perform better under incentive contracts than under non-contingent payment systems. Asch (1990) shows that Navy recruiters vary their effort over time in response to piece rates, quotas, and prizes. This literature, which is more expansive than the studies mentioned here, indicates a considerable effect of compensation on worker behavior.

The empirical work in this paper contributes to both strains of the literature: first, the variation of observed contracts, and second, the extent to which incentives affect behavior. The contract section uses very detailed firm-level data and tests a specific proposition from agency theory, directly addressing Prendergast's 'theoretical identification' issue. Since the contract data is only available for a cross section of firms, though, I cannot identify *causality* of monitoring on contract choice. But the results do speak to a more general prediction, as in Holmstrom and Milgrom (1994), of a *complementarity* between monitoring and incentive pay. The effect of incentives on employee effort is able to be more precisely addressed due to the existence of a time series dimension of the data. I am able to interpret the results as identifying a causal impact of monitoring on variables that represent direct measures of driver behavior (i.e. truck life expectancy and fuel efficiency).

It's important to note that a substantial recent literature has emerged that uses the trucking industry to examine contracting and other organizational issues. A few of these papers study issues similar to the ones addressed here. Belman and Monaco (2001) document positive effects of OBCs on drivers' earnings, consistent with improved efficiency and incentives. Lafontaine and Masten (2002) argue that driver contracts are structured to minimize the costs incurred in price determination for heterogeneous hauls. Finally, Baker and Hubbard (2001) look at the impact of the incentive capabilities of OBCs on whether shippers use trucks from their own fleet for a haul or employ a for-hire trucking firm.

# 3 The Trucking Industry and On-Board Computers

Since the end of federal regulation in the early 1980's, the trucking market has been extremely competitive. Trucking firms, called 'carriers', differentiate themselves to their customers, called 'shippers', along several different dimensions.[5] These dimensions include length of haul (local, short-range, medium-range, and long-range), type of product carried (for example, chemicals and petroleum, refrigerated products, bulk materials, and dry cargo), size of the shipment, and the length of the service contract, which varies from spot-market arrangements, called common carriage, to longer term contracts between carriers and shippers called contract carriage.[6] Carriers employ two types of drivers. 'Owner-operators', as the name implies, own their trucks, and lease their driving services to carriers on a haul by haul basis. 'Company drivers' are employees of the trucking firm, and are paid to ship products using the firm's trucks and equipment. As is discussed below, which party owns the truck has tremendous implications for how incentives are structured. Truckers are predominantly company drivers; only 10% of trucks are driven by owner-operators. The analysis in this paper focuses exclusively on company drivers, and studies how the introduction of the OBC technology changes the nature of their relationship with the carriers who employ them.

OBC technology began to be implemented in the trucking industry in the late 1980's, carrying with it expectations of immediate improvements in the efficiency of operations. There are two classes of OBCs: a more primitive version, called a trip recorder, and a more sophisticated device, called an Electronic Vehicle Management System (EVMS). Trip

[5]There are two types of carriers. 'Private fleets' are subsidiaries of non-trucking firms, and are primarily used to ship their own products. 'For-hire' carriers exist solely to ship the cargo of other parties.

[6]The size of shipment market is divided into truckload (TL) and less-than-truckload (LTL) carriers. The TL sector consists of very large shipments, where a haul typically contains cargo from a single shipper. The LTL sector combines smaller shipments from several sources.

recorders, introduced slightly earlier than EVMS, keep an electronic summary of how a driver operates a truck. A trip recorder is activated once a driver begins a haul, and the contents become accessible only when the driver returns back to his carrier. At this time, the data can be downloaded to computers at the firm, processed using standard industry software, and then analyzed by the driver's superiors. The information collected by a trip recorder includes departure and arrival times, revolutions per minute of the engine, periods of stop-and-go driving, brake use, and precise measures of fuel consumption. Also recorded are the three most important forms of driver behavior that wear down the operating value of the truck: excessive speed, idling time, and over-revving of the engine. While this information is valuable for mechanics who may need to diagnose engine problems, it is primarily useful because it allows firms to discern exactly how drivers drive trucks. Firms are willing to pay for this knowledge since both a truck's value and its fuel efficiency are very sensitive to how the truck is operated. With the incentive-enhancing information provided by trip recorders in hand, firms can potentially better shape driver behavior through more efficient contracting. The cost to purchase and install a trip recorder was about $500 in the early 1990's, with this price remaining relatively constant through the end of the decade.

EVMS provide firms with all of the information-collecting capabilities of trip recorders, as well as several extra features. These additional features include real-time communication between drivers and carriers through e-mail type messaging, knowledge of the exact location of trucks via global positioning satellite technology, and the ability to provide real-time tracking information to customers on the location of their shipments. These attributes of EVMS, in particular the instantaneous data received from trucks and the facile means of communication, can improve resource allocation decisions within the firm. Several papers have studied the EVMS technology in this regard.[7] Anecdotally, there is no evidence that

[7]See Hubbard (2000, 2001) and Cacciola (2002).

the real-time *incentive* features of EVMS are used by carriers.[8] Therefore, the contracting improvements possible with trip recorders and EVMS are identical, and no distinction is made between the two technologies in the analysis that follows. The cost of purchasing and installing EVMS is substantially larger than that of a trip recorder. In 1997, a single terminal cost between $2,500 and $4,000, with monthly communication fees of $50 to $100 per truck. Additional installation costs can also run a few thousand dollars.

# 4 A Model of Worker Incentives and Contract Choice

This section presents a model that yields several testable implications that are addressed in the empirical results to follow. A Holmstrom-Milgrom (1991) multitask principal-agent model of contract choice is used to describe the carrier and company driver relationship. As is customary, the model is represented as a Stackelberg game with two players; here, the parties are a trucking firm (the principal) and a driver (the agent). In the first stage, the firm selects the terms of a compensation contract to offer the driver, contingent on having chosen to adopt (or not adopt) an OBC. In the second stage, the driver decides whether to accept or reject the contract, and, conditional on accepting, chooses his effort levels.

The substance of a truck driver's job is to move cargo from one location to another. An additional characteristic of the job that is significant to the carrier is the *manner* in which cargo is transported by the driver. As such, an accurate description of the driver's role in production must consider this dual nature of the driver's efforts. To capture this element of reality, I consider a model where the driver's job consists of two tasks: (i) to transport the product in a timely fashion from one location to another (the 'productivity' task), and

[8] For example, supervisors at a carrier do not use EVMS to see if a driver is speeding at a given point in time, and then immediately tell him to slow down. Communication is limited to issues regarding the scheduling of shipments.

(ii) to drive the truck in a manner that is not abusive and maintains the truck's value (the 'operation' task). It is obvious from considerations of reputation and customer service that a carrier places great importance on the consistent completion of the productivity task. But for several reasons the carrier also has considerable interest in the effort directed towards the operation task. First, poor driving technique, as characterized by driving at high rates of speed, accelerating quickly, shifting erratically, idling excessively, and over-revving the engine, stresses the mechanical structure of the truck, potentially causing part failures and reducing truck life expectancy. Second, this type of behavior lowers fuel efficiency, which is of interest to the carrier since it is the party that pays for fuel expenses. Third, high speeds increase the probability of an accident, which may damage the truck, shipment, and driver.

The driver, however, may prefer to drive the truck in a way that is not desirable to the carrier. For example, maintaining a higher average speed while on the road allows the driver to take longer breaks, and yet reach his destination on time. Also, a driver can often change the operating specifications of the truck to suit his own desires, sometimes at the expense of the carrier. One driver describes a particular form of sabotage at a firm where he worked: "Some drivers secretly altered their trucks' fuel pumps to increase the engine's horsepower, a practice know as 'jacking up.' A jacked-up pump is likely to cost the owner by cutting fuel mileage, lowering the engine's life expectancy, and putting more wear on the drive train and drive tires."[9] The divergence of driver and carrier preferences in terms of driving technique is at the crux of the contracting decision, particularly when driver behavior can be difficult to observe.

In terms of notation, let the amount of agent effort used in the productivity task and the operation task be $t_1$ and $t_2$, respectively, so that the driver's job is the vector $t = (t_1, t_2)$. Assume that the personal cost of effort to the agent is $C(t)$, and that this is a strictly

[9]Quoted from Ouellet (1994: 85-86).

convex function. The firm's gross benefit, which depends on agent effort, is labeled $B(t)$, and is assumed to be strictly concave in its arguments. I assume further that the outputs associated with $t_1$ and $t_2$ are noisy versions of true driver effort. In particular, the firm observes the vector of outputs influenced by the driver, $y = (y_1, y_2)$, where $y_j = t_j + \epsilon_j$ for $j = 1, 2$, and $\epsilon_1$ and $\epsilon_2$ are random variables with mean 0 and variances $\sigma_1^2$ and $\sigma_2^2$, respectively, and covariance $\sigma_{12}$.

The impact of OBC adoption is manifested in how accurately driver efforts are measured, and are thus reflected in the variances in output, $\sigma_1^2$ and $\sigma_2^2$. The productivity task, $t_1$, is nearly perfectly observable by the carrier at little cost, even without an OBC. Late arrivals or damaged products are generally reported by the shipper to the carrier. Also, factors outside of the driver's control that may cause delays, such as traffic and weather conditions, are easily verifiable. Thus, $\sigma_1^2$ is relatively low, and falls only trivially with use of an OBC. The amount of effort directed towards the operation task, $t_2$, on the other hand, is extremely difficult to measure without an OBC. Measures of this task are very noisy, in part because carriers cannot easily distinguish between mechanical problems caused by bad driving and those associated with the normal wear and tear of truck use. This is compounded when shipment schedules and truck assignments dictate that more than one driver use the same truck for different hauls. One obvious way to resolve this problem is to give the ownership rights of the truck to the driver. In this case, the driver will internalize the costs associated with poor driving behavior.[10] The operation task becomes *significantly* better measured if

[10]Perhaps somewhat surprisingly, only a small fraction of drivers are owner-operators (on the order of 10%). One possible explanation is that driver ownership improves incentives *within* hauls, but worsens incentives *between* hauls, since owner-operators have the incentive to search for alternative hauls that pay a higher rate for their services (see Baker and Hubbard (2000)). Other explanations include the notion that certain types of service require significant levels of coordination, and that carriers need to invest in a reputation with shippers and thus will be more likely to rely on their own drivers (see Nickerson and

a truck is equipped with an OBC, and is represented in the model by a large decrease in $\sigma_2^2$ upon adoption.

To proceed with the model, I assume that the optimal contract is linear, and is of the form $w(t) = w_0 + a_1 y_1 + a_2 y_2$, where $w_0$ is the fixed component of the contract and $a_1$ and $a_2$ are the 'incentive intensities', which tie pay to measured driver performance.[11] Assume also that the driver has a utility function given by $U(x) = e^{-rx}$, where $r$ is the driver's constant coefficient of absolute risk aversion. The principal is assumed to be risk neutral. Given the linear contract and the production function specified above, the agent's net payoff is:

$$w_0 + a_1 t_1 + a_2 t_2 - C(t) + a_1 \epsilon_1 + a_2 \epsilon_2 \tag{1}$$

Under the assumed utility function, the agent's certainty equivalent (CE) is:

$$CE = w_0 + a_1 t_1 + a_2 t_2 - C(t) - \frac{1}{2} r(a_1^2 \sigma_1^2 + 2a_1 a_2 \sigma_{12} + a_2^2 \sigma_2^2) \tag{2}$$

where the last term is the agent's risk premium. The principal's expected payoff is the gross benefit minus the expected cost of compensation and the cost of investing in an OBC ($I$):

$$B(t) - w_0 - a_1 t_1 - a_2 t_2 - I \tag{3}$$

So adding equations (2) and (3) together, the total certainty equivalent (the joint surplus) is:

$$B(t) - C(t) - I - \frac{1}{2} r(a_1^2 \sigma_1^2 + 2a_1 a_2 \sigma_{12} + a_2^2 \sigma_2^2) \tag{4}$$

Silverman (1999)).

[11] Mirrlees (1974) shows that a nonlinear step-function can actually outperform the best linear contract in a one-shot game. Holmstrom and Milgrom (1987) reinterpret the model to include a sequence of actions by the agent which yields a sequence of outcomes, and show that the optimal contract is actually linear in the aggregate output. Moreover, their paper illustrates that linear contracts are more robust to 'gaming' by the agent than are nonlinear contracts.

Notice that the fixed payment $w_0$ drops out of this expression. This component serves only to divide the joint surplus between the two parties, and to ensure that the participation constraint needed to induce the agent to take the job is satisfied.

In terms of choosing a monitoring technology, the firm has three choices: (i) don't adopt either technology, (ii) adopt a trip recorder, or (iii) adopt an EVMS.[12] Determining the socially optimal adoption of technology simply requires a comparison of the joint surpluses under each of the three options. In lieu of developing a full-scale model of adoption, a few notes on the adoption decision should suffice. First, and most obviously, adoption ultimately is a cost-benefit analysis. Firms compare the perceived benefits of adoption resulting from improved incentives with the costs of purchasing, installing, and utilizing the OBC systems. Second, and most important from an empirical point of view, variation in adoption is driven by the fact that the benefits of the technology vary by sector of the trucking industry. For example, consider a comparison of short-haul trucks with long-haul trucks. Trucks that operate close to their home base and make regular deliveries can be monitored by other means besides OBCs. CB radios can be used to check the status of trucks, and the fact that these drives have regularly scheduled drop-offs and pick-ups means that there is little scope to drive according to one's own personal preference. Long-haul truckers, on the other hand, are out of range of CB radios, and, since they are on the road for days at a time, have tremendous latitude in how they schedule their time. The incentive benefits of OBCs are thus much greater for long-haul than short-haul truckers, and we would expect to see adoption rates reflect this fact.

To continue with the model, recall that the OBC adoption decision is manifested in

[12]While this paper only considers the incentive effects of OBCs, which are identical for trip recorders and EVMS, a more general adoption decision for a firm considers the resource allocation capabilities of EVMS. This additional feature of EVMS will in part drive the variation that we see empirically, but it is not discussed here.

the variance of the output parameters, $\sigma_1^2$ and $\sigma_2^2$. Therefore, the optimal contract is conditional on OBC adoption; adoption and non-adoption imply different values for the variance parameters, and will thus cause different choices for the endogenous variables $a$ and $t$. The optimal linear contract is the one that maximizes the joint surplus in equation (4) with respect to the choice of $t$ by the agent and $a$ by the principal, subject to the agent's incentive compatibility (IC) constraint. Formally, the optimal contract is characterized by $(w_0, a, t)$ and solves:

$$\max_{(a,t)} B(t) - C(t) - I - \frac{1}{2}r(a_1^2\sigma_1^2 + 2a_1a_2\sigma_{12} + a_2^2\sigma_2^2) \tag{5}$$

subject to

$$\max_{t'} a_1t_1' + a_2t_2' - C(t')$$

Assuming a strictly interior solution for the agent's IC constraint, we can rewrite it in terms of its first order conditions: $a_1 = C_1(t)$ and $a_2 = C_2(t)$, where $C_j(t)$ indicates the partial derivative with respect to the $j$th argument of $C(t)$. It is important to see that these first order conditions indicate the responsiveness of agent effort, $t$, to changes in the compensation parameters, $a$. By differentiating, it can be shown, for example, that

$$\frac{\partial t_1}{\partial a_1} = \frac{C_{22}}{D} \text{ and } \frac{\partial t_1}{\partial a_2} = -\frac{C_{12}}{D} \tag{6}$$

where $D$ is the determinant of $\frac{\partial^2 C}{\partial t^2}$ and is positive. Therefore, the agent's effort towards $t_1$ increases with $a_1$ (since $C_{22}$ is positive), and decreases with $a_2$ if $C_{12} > 0$. The cross-partial derivative, $C_{12}$, indicates the complementarity or substitutability of the two tasks in the agent's cost function. In the trucking context, the tasks are substitutes ($C_{12} > 0$); the tension between the productivity task and the operation task means that directing greater effort towards one task results in a greater marginal cost of performing the other task. As the equations in (6) indicate, increasing the incentive intensity on one task not only affects

the attention to that particular task, but also influences the effort directed towards the other task. This effect of the incentive intensities on allocating effort across tasks (in addition to their role in allocating risk and motivating hard work as in the single task principal-agent model) is key to understanding the optimal contract.

Now, the optimal incentive intensities, $a_1$ and $a_2$, can be determined. I will assume that the complete output vector $y = (y_1, y_2)$ is observable (i.e. both $\sigma_1^2$ and $\sigma_2^2$ are finite), and so the optimal contract will depend on both $a_1$ and $a_2$. I also assume that there is zero correlation across the noise terms, so that $\sigma_{12} = 0$. Under these conditions, the optimal incentive intensity for $t_1$ is

$$a_1 = \frac{B_1(1 + \frac{1}{rC_{22}\sigma_2^2}) - \frac{B_2C_{12}}{C_{22}}}{1 + \frac{1}{rC_{22}\sigma_2^2} + r\sigma_1^2(C_{11} - \frac{C_{12}^2}{C_{22}}) + \frac{C_{11}\sigma_1^2}{C_{22}\sigma_2^2}} \tag{7}$$

and the optimal incentive intensity for $t_2$ is

$$a_2 = \frac{B_2(1 + \frac{1}{rC_{11}\sigma_1^2}) - \frac{B_1C_{12}}{C_{11}}}{1 + \frac{1}{rC_{11}\sigma_1^2} + r\sigma_2^2(C_{22} - \frac{C_{12}^2}{C_{11}}) + \frac{C_{22}\sigma_2^2}{C_{11}\sigma_1^2}} \tag{8}$$

In order to interpret these formulas, a few facts are noteworthy. First, the denominators in both expressions are strictly positive due to the assumed strict convexity of the agent's cost function. Second, the $C_{12}$ term and the precision in measuring effort, $\sigma_1^2$ and $\sigma_2^2$, play an important role. The trucking context is characterized by both the substitutability of the tasks ($C_{12} > 0$), and the difficulty in measuring the operation task, $t_2$, without use of an OBC (indicating that $\sigma_2^2$ is large). In this case, the incentive intensity on the productivity task is muted, as the firm fears that rewarding this task will result in the neglect of the operation task.

Adoption of an OBC causes a significant reduction in the $\sigma_2^2$ term, and thus a reweighting of $a_1$ and $a_2$. An examination of the comparative statics of the incentive intensities with

respect to a change in $\sigma_2^2$ provides precise and intuitive predictions:[13]

$$\frac{\partial a_2}{\partial \sigma_2^2} < 0 \text{ if } a_2 > 0 \tag{9}$$

and

$$\frac{\partial a_1}{\partial \sigma_2^2} < 0 \text{ if } a_2 > 0 \text{ and } C_{12} > 0 \tag{10}$$

So conditional on the firm providing a positive incentive intensity for the second task, adoption of an OBC and the decrease in $\sigma_2^2$ will unambiguously increase $a_2$, resulting in stronger incentives for the operation task. Again conditional on $a_2$ being positive, when $C_{12}$ is positive, the decrease in $\sigma_2^2$ will increase $a_1$, so that the productivity task is also weighted more heavily than prior to adoption.[14] In sum, under adoption of an OBC we expect to see higher powered incentives for *both* tasks. Data is brought to bear on this prediction below.

# 5 On-Board Computer Adoption and Contract Variation

The implementation of OBCs is a unique situation that provides a direct and observable change in one of the parameters that influences the optimal compensation contract. The comparative static results in (9) and (10) predict that a decrease in $\sigma_2^2$, as embodied by

[13]The exact formulas for the comparative statics are

$$\frac{\partial a_2}{\partial \sigma_2^2} = -\frac{[B_2(1+\frac{1}{rC_{11}\sigma_1^2}) - \frac{B_1 C_{12}}{C_{11}}][r(C_{22} - \frac{C_{12}^2}{C_{11}}) + \frac{C_{22}}{C_{11}\sigma_1^2}]}{[1+\frac{1}{rC_{11}\sigma_1^2} + r\sigma_2^2(C_{22} - \frac{C_{12}^2}{C_{11}}) + \frac{C_{22}\sigma_2^2}{C_{11}\sigma_1^2}]^2}$$

and

$$\frac{\partial a_1}{\partial \sigma_2^2} = -\frac{[B_2(1+\frac{1}{rC_{11}\sigma_1^2}) - \frac{B_1 C_{12}}{C_{11}}][\frac{C_{11}C_{12}\sigma_1^2}{C_{22}^2(\sigma_2^2)^2}]}{[1+\frac{1}{rC_{22}\sigma_2^2} + r\sigma_1^2(C_{11} - \frac{C_{12}^2}{C_{22}}) + \frac{C_{11}\sigma_1^2}{C_{22}\sigma_2^2}]^2}$$

[14]It is interesting to note that when the tasks are complementary ($C_{12}$ is negative), that a decrease in $\sigma_2^2$ causes a decrease in $a_1$. This is because with the increased precision in measuring $t_2$ and the corresponding increases in $a_2$ and $t_2$, complementarity implies that the marginal cost of $t_1$ will fall. So the agent will direct *more* effort in this direction, and in fact the principal needs to place a brake on the effort directed toward $t_1$ by reducing $a_1$.

adoption of an OBC, causes an increase in the incentive intensities for both the productivity and operation tasks. The empirical work in this section addresses this prediction by analyzing how OBC adoption and incentive pay covary across sectors of the trucking industry.

## 5.1 Data

There are two sources of data used in this section. The first is the Census of Transportation's Vehicle Inventory and Use Survey (VIUS) for the 1997 survey year.[15] The VIUS provides data on physical and operating characteristics for a random sample of the U.S. truck population. This paper uses the observations on truck-tractors, which are the front-end power-units of the truck-and-trailer combinations. In the 1997 survey, there are 25,533 observations (truck-tractors). Key variables included in the VIUS are trip recorder and EVMS use, length of haul, type of trailer attached, principal products hauled, fleet size, model year, and average miles per gallon.

The second source of data is the National Survey of Driver Wages, a firm-level data set containing information on driver base pay and bonuses of for-hire carriers operating in the truckload sector.[16] In the August 1999 data set there are 241 firms surveyed that employ company drivers. Drivers' base pay is almost exclusively determined on a per mile basis. A small number of firms pay based on 'percentage', which means that a driver's pay is determined as a fraction of the shipment's revenue to the carrier. In addition to this base pay, drivers are often compensated with bonuses if they achieve certain targets. This survey includes information on the intensities of four bonus categories. A 'productivity' bonus is awarded to drivers who obtain a mileage target set by the firm. A 'performance' bonus is given to drivers for on-time deliveries and the prompt completion of truck logs and

[15]Previous surveys were called the Truck Inventory and Use Survey (TIUS).

[16]This data is collected by a private firm called Signpost Inc.

paperwork. These two bonuses are in the spirit of rewarding the productivity task, $t_1$, as they reflect a driver's effort in transporting cargo promptly for shippers. Two additional bonuses are more in the vein of how a driver actually operates a truck, and are indicative of the driver's effort towards the operation task, $t_2$. A 'fuel efficiency' bonus is awarded to drivers based on idling time, average speed, and miles per gallon. Finally, a 'safety' bonus is paid based on accident-free miles.

## 5.2 Empirical Results

Table 1 displays OBC use rates by two central dimensions of the trucking industry, length of haul and the type of trailer attached to the truck-tractor. In the full sample, about 18% of trucks surveyed in 1992 were outfitted with OBCs, with this use rate increasing to nearly one-third of trucks by 1997. Within both the length of haul and trailer type categories there is substantial variation in adoption rates. OBC use increases monotonically with length of haul, topping out at over 50% in 1997 for long range trucks making hauls beyond 500 miles. This phenomenon, alluded to earlier, reflects the fact that long-haul truckers have more latitude in how they operate their trucks, and so the benefit of monitoring is greater for this group of drivers. There are large differences in adoption across trailer types as well. Refrigerated vans have particularly high use rates of OBCs, consistent with the notion that the value of having a record of driver behavior is large when late arrivals are costly. The variation in adoption rates across sectors is obviously crucial to estimating the effect of the technology on industry variables.

Table 2 provides summary statistics of the number of firms offering particular types of bonuses and the metric used to pay them (i.e. per mile, lump sum, or percentage), as well as the mean bonus payment and the mean base pay rate for each of the payment categories. A safety bonus is the most common, with just under 50% of firms offering one, while the other

types of bonuses have incidences between 20% and 30%. Most of the bonuses are offered on a per mile basis, though some firms offer lump sum payments for achieving targets, most notably for the safety bonus. The bonus amounts are non-trivial: for those firms paying bonuses on a mileage basis, the mean bonus payment as a percentage of the mean base pay rate ranges between 5% (safety) and 8% (productivity). It's also interesting to note that for those firms that pay bonuses per mile, their mean base pay rate is lower than the mean base pay rate for firms not offering bonuses. But if a driver achieves the target and the bonus is paid out, then total compensation (base pay rate plus bonus payment) for the firms offering bonuses exceeds total compensation for those not offering bonuses. Moreover, this intuitive result is true for all four bonus categories.

In order to match the OBC data (from 1997) with the compensation data (from 1999), I create cells on the basis of base state, cargo type, and length of haul.[17] Cell-level averages are then computed for each of the relevant variables. This results in a single cross-section of cell-level data with 112 observations. In computing the cell-level averages, for the bonus payments I only use information on whether or not a firm offers a particular bonus, and do not include the level that it offers. There are two reasons for doing so. First, this allows for a pooling of bonuses across payment types (i.e. mileage, lump sum, and percentage). Second, conditional on offering a bonus, there is little variation in the level of the payment, so this simplification likely throws away no meaningful information.

The correlation coefficients between driver bonuses and OBC use are documented in Table 3. Of primary relevance to the testing of agency theory is the lower-left quadrant of the table. Three of the four correlations relating OBC use and bonuses are positive, including both of the bonuses reflecting the operation task. The fuel efficiency bonus and

[17]To make the data from the VIUS comparable to the sample of firms in the compensation data, I limit the trucks used in the VIUS to those driven by company drivers employed in for-hire firms in the TL sector.

OBC variable have a particularly large correlation coefficient of .222. Table 4 attempts to make more precise the statistical relationship between the variables by providing regressions of each bonus variable on the OBC use rate, as well as controls for length of haul, cargo type, and base state. Both unweighted and weighted estimates are displayed, where the weighted regressions are weighted by the sum of the number of firms from the compensation data and the number of trucks from the OBC data in a given cell. The weighting scheme used reflects the fact that some cells represent a larger proportion of the industry than others, and so should be given a correspondingly greater influence on the results. Both sets of regression estimates qualitatively reproduce the simple correlation coefficients. The two bonuses associated with the productivity task, the productivity bonus and the performance bonus, display a near zero association with OBC use. The operation task, though, which becomes more precisely measured upon adoption, does appear to have bonuses that are positively associated with OBC use. The fuel efficiency and safety bonuses have fairly large, positive point estimates for the OBC variable. The OBC coefficient in the fuel efficiency regression, while attenuated a bit in the weighted version, is significant at the 5% level in the unweighted regression. The point estimates in both the fuel efficiency and safety bonuses center around .2, indicating that firms that outfit their entire fleet of trucks with OBCs are 20% more likely to offer these bonuses than firms with a zero adoption rate.

Overall, the test of the agency theory predictions regarding contract choice provides mixed results. Theoretically, the bonus rates for both tasks are predicted to be greater under OBC adoption, but evidence is found that only bonuses for the task that is more precisely measured are more likely. It's clear that power is an issue here, as 112 observations is inadequate for delivering precise estimates. Also, this cross-sectional data does not allow inference as to the causal impact of OBC use on contract choice, but the evidence is supportive of the proposition that monitoring and incentive pay covary positively in the

data. This is consistent with more general models of firm behavior, such as Holmstrom and Milgrom (1994), where monitoring and incentive pay are complementary instruments for motivating workers.

It is also important to note that while bonuses are a key part of the observed contract, there are several other more indirect components, unobserved in the data, that are given to reward or punish driver behavior. Managerial decisions that affect job and route assignments, the schedule of working hours, and the matching of drivers to particular trucks all directly impact a driver's utility. Certainly these instruments can be used by firms, as well as the direct monetary components of incentive pay, to motivate driver effort.[18]

# 6 How Do Workers Respond to Changes in the Monitoring Environment?

The adoption of an OBC unambiguously changes a carrier's ability to monitor driver behavior, and the previous section provides some suggestive evidence that the nature of the compensation contract changes correspondingly. This section estimates the magnitude of the incentive effect by analyzing how drivers alter their behavior upon adoption of an OBC. The data allows for detection of an effect regarding two observable variables: truck life expectancy (truck age) and fuel efficiency.

## 6.1 Data

This section uses the Census' TIUS data for the 1987 and 1992 surveys, and the VIUS for the 1997 survey. OBCs were not available as of the 1987 survey, so adoption begins to emerge in 1992. Other important variables not already discussed include a set of variables

[18]See Ouellet (1994) for a comprehensive discussion of the inner workings of the trucking industry from a driver's perspective.

recording truck accidents in the 1987 survey. Also note that the surveys do not form a panel; different trucks are included in each data set.

## 6.2 The Effect of On-Board Computers on Truck Age

Since the monitoring capabilities of OBCs allow firms to better understand how drivers operate trucks, carriers can potentially devise means of improving driver effort directed towards the operation task, $t_2$. Drivers can be instructed as to how to improve their driving technique, resulting in less wear and tear on the truck. This suggests that adoption of an OBC should increase a truck's life expectancy. The truck-level behavioral equation to be estimated is the following:

$$AGE_{i,t+1} = \beta OBC_{it} + X'_{i,t+1}\gamma + \epsilon_{i,t+1} \tag{11}$$

where $i$ indexes trucks and $t$ indexes time; $AGE$ is truck age; $OBC$ equals one if an OBC is installed on the truck and zero otherwise; $X$ is a vector of control variables that influence truck age, for example, length of haul, trailer type, principal products hauled, and fleet size; $\epsilon_{i,t+1}$ is a white-noise error; and $\beta$ and $\gamma$ are parameters. The parameter $\beta$ is hypothesized to be positive, indicating that OBC use should extend truck age compared with non-adoption, everything else held constant. It is important to note the timing of the variables in the equation. OBC adoption at time $t$ will only result in a detectable effect on truck age after the normal life-cycle of the truck passes. Thus, the dependent variable, truck age, is specified at time $t + 1$. Summaries of the truck age variable indicate that many trucks have a long life span. A substantial fraction of trucks, on the order of 30% in each of the three surveys, are ten years old or greater. This emphasizes the fact that a suitable amount of time must pass after OBC adoption in order to detect improvements in truck life expectancy.

The difficulty in attempting to identify a causal impact of OBC use on truck age is that newer trucks are more likely than older trucks to have OBCs, simply because new trucks

tend to come bundled with the latest technology.[19] This statement does not reflect any behavioral content regarding driver behavior, but simply captures a technological change in the industry. This more mechanical relationship is given by:

$$OBC_{it} = \delta AGE_{it} + Z'_{it}\upsilon + \upsilon_{it} \tag{12}$$

where $OBC$ and $AGE$ are as defined above; $Z$ is a vector of control variables that influence OBC adoption, and include length of haul, principal products hauled, trailer type, and accident rates; $\upsilon_{it}$ is a white-noise error; and $\delta$ and $\upsilon$ are parameters. The parameter $\delta$ is hypothesized to be negative, reflecting the fact that older trucks are less likely to have OBCs. Table 5 explores this notion, and the results are quite striking. Linear probability model estimates are provided where the dependent variable is OBC use. The covariates of interest are a set of truck age dummy variables, ranging from one year old to ten years and older.[20] For both the 1992 and 1997 surveys, OBC use substantially decreases as truck age increases. The magnitudes are large, particularly in the 1997 survey, where a ten year old truck is about 42% less likely than a new truck to have an OBC installed. These results indicate a considerable hurdle that must be overcome in order to identify causality running in the other direction: that is, from OBC use to truck age.

The timing issue of equations (11) and (12) is connected by the sector specific replacement rates of trucks. Some sectors of the industry replace their trucks faster than others, independent of the survey year. For example, long-haul trucks typically drive more miles per year than local trucks, reducing the amount of time that they can be used effectively on the road. Consequently, long-haul trucks must be replaced more frequently, and overall

[19] Cacciola (2002) discusses this phenomenon in detail.

[20] Control variables include length of haul, trailer type, TL/LTL, private fleet and contract carriage, fleet size, base state, principal product hauled, intrastate operation, owner-operator, private refuel facility (as opposed to refueling at truck stops), and exempt carrier (a vestige of regulation).

are of a younger vintage than local trucks. This phenomenon is true across survey years, inducing a positive association between truck age at time $t$ and truck age at time $t+1$ across sectors. Of course, if there is a positive behavioral impact of the technology, then this positive association should dampen over time. The younger sectors of the industry, for example the long-haul trucks, because they replace their trucks more frequently are more likely to adopt OBCs due to the diffusion of the technology over time. If the behavioral component is present, then long-haul trucks should last longer than prior to adoption, moving their age distribution closer to that of the low-adopting and older vintage local trucks.

The above exposition makes clear that a time series dimension to the data is necessary. Since the surveys themselves do not form a pure panel of trucks, I create a synthetic panel data set. Trucks are divided into cells defined by base state of operation, trailer type, principal product hauled, and length of haul, and then cell-level averages are computed. Cells can then be tracked over time, mimicking the nature of a true panel data set. Deaton (1985) emphasizes that the cell-level *sample* averages of the variables in the data are error-ridden proxies of the true *population* means. Therefore, measurement error in the independent variables may bias certain estimators away from their intended targets.

Turning to the empirical results, Table 6 shows OLS regressions of the relationship between truck age and OBC use. The first column is a simple cross-section regression of truck age in 1992 on OBC use in 1992, including several control variables. The large and significant negative coefficient is a manifestation of equation (12) above. Older trucks are less likely to have OBCs, simply because the newer trucks tend to come equipped with OBCs as part of the standard option package. Also, the behavioral impact of equation (11) has not yet had time to percolate, since the regression consists of contemporaneous measures of truck age and OBC use. Thus, it is not at all surprising that the coefficient is so strongly negative. In the second column, the dependent variable is truck age in 1997, while the OBC

variable is measured in 1992. In this case, enough time has passed for the behavioral impact to express itself, but it is still overshadowed by the mechanical effect of equation (12). The fact that the coefficient has attenuated down closer to zero is some evidence, though, that the behavioral impact is present in the data. In other words, the first column coefficient contains the mechanical effect, while the second column coefficient embodies both the mechanical and behavioral effects. So the increase in the coefficient from column 1 to column 2, in a loose sense, is indicative of the actual causal effect of OBC use on driver behavior.

Table 7 presents two direct approaches of correcting for the endogeneity of OBC adoption. The first column uses age *growth* from 1992 to 1997 as the dependent variable. This effectively conditions on the past truck age distribution by sector, removing the differential OBC adoption rates due to differing truck replacement rates. The coefficient is large and significantly positive, indicating that OBC use extends truck life expectancy by just under on year. The framework of equations (11) and (12) also naturally suggest an instrumental variables strategy. The key is to find some variables $Z'_{it}$ that affect OBC adoption, but not truck replacement. Ideal candidates for instruments are variables that isolate the variation in OBC adoption due solely to incentives. In the 1987 survey, there is information on four accident variables: a general accident indicator, whether the accident involved a fatality, whether the accident involved bodily injury needing medical treatment, and whether there was property damage in excess of $4,200. The excluded instruments that I use are these accident variables as well as these variables interacted with the product type dummy variables. Conceptually, the idea is that cells with high accident rates value OBCs' incentive benefits in order to better monitor their drivers, as well as have a record of driver behavior to provide insurance companies. The interactions are intended to capture the fact that product types are of varying monetary value, so that cells with relatively large accident rates carrying highly priced shipments are in particular need of OBCs.

Appendix Table 1 provides the results of the first stage regression, where OBC use in 1992 is the dependent variable. Three sets of covariates are listed: the four accident variables, the product type main effects, and the interactions of the accident rates with the product dummies.[21] Cells in the chemical product category are the omitted group for all sets of covariates. Therefore, all of the interactions are in comparison to the effects of accident rates on OBC use for the chemical group. Interaction terms that are positive (negative) indicate the product groups that are more (less) responsive to accident rates than the chemical product category in terms of OBC adoption. The chemical category is chosen as a basis for comparison because it has relatively highly valued shipments and the additional consequences of crashes can be quite large. The coefficients for the accident variables provide some justification for the identification concept described in the previous paragraph. Three of the accident variable coefficients for the chemical category (the number of fatalities, the number of bodily injuries, and the number of accidents causing property damage greater than $4,200) are positive, as expected. Somewhat surprisingly, the number of accidents has a negative coefficient. A piece of evidence consistent with the identification scheme is that nearly all product categories are less responsive to three of the accident variables in terms of adoption than is the chemical category. Again, the exception is the interactions for the total number of accidents, where most product categories are more responsive than the chemical group. Finding a reasonable pattern amongst the other product categories is somewhat difficult. Textile mill products, which tend to be expensive shipments, do display coefficients very similar to those from chemical products (in fact, none of the four accident variable coefficients for textile products is significantly different from the chemical

[21]Note, though, that the product type main effects are *not* excluded instruments; they are included in both the first and second stages. The reason for this is that product type is a determinant of truck replacement rates, and thus truck age.

products). Overall, the accident variables and the interaction terms have very good power in predicting OBC adoption; the 73 excluded instruments are jointly significant for a test of size 0.0000.

There is some question as to whether these variables pass the exclusion restriction requiring that they not be present in the second stage equation, in that accidents damage trucks and effectively reduce truck age. But since accident rates are extremely low, their effect on truck age is probably minimal. Accidents are primarily costly along other dimensions, such as damaged cargo, decreased goodwill with shippers, higher insurance costs, and a loss of reputation. It is these costs of accidents that primarily drive firms to use OBCs for incentive purposes. In any event, the IV model is heavily over-identified, so the null hypothesis that the excluded instruments are valid can be tested. The over-id test has a statistic of $\chi^2(72) = 175$, which yields a p-value of 0.0000, strongly rejecting the null hypothesis. The IV results, displayed in the second column of Table 7, should be interpreted with this caveat in mind. The second stage coefficient on OBC use is positive and significant at the 10% level. The point estimate of the coefficient is very close to the coefficient in the first column that is obtained using age growth.[22] Taken together, the results in Tables 6 and 7 provide strong evidence that the adoption of an OBC changes driver behavior in ways that prolong a truck's lifetime.

## 6.3 The Effect of On-Board Computers on Fuel Efficiency

A second observable variable with which we can attempt to detect changes in driver behavior is fuel efficiency (or miles per gallon (mpg)). Recall that quick accelerations and driving at

[22] As a comparison, an IV model that only includes the four accident variables, and drops the interactions altogether, has much less power than one that includes the interaction terms. The coefficient on OBC use in the former model is 3.47 with a standard error of 3.45.

high rates of speed reduce fuel efficiency. Since carriers pay for gas, they have an incentive to monitor driver behavior along these dimensions, and OBCs allow them to do so. It is useful to discuss how an OBC is expected to improve fuel efficiency. Even without an OBC, rough estimates of mpg can be estimated by simply dividing length of haul by the gallons of fuel consumed. Nonetheless, the use of an OBC provides slightly more precise estimates of true mpg. But the main contribution of an OBC is that it details exactly *how* a particular level of fuel efficiency is achieved. Without an OBC, a driver can blame subpar mpg on several factors, such as poor engine performance or traffic delays that interfere with efficient operation of the truck. But with an OBC, the *exact* reasons for a particular mpg outcome are apparent, including driver behavior. Trucking firms can also use their analyses of the OBCs' contents to provide drivers with instructions as how best to improve their driving technique.

The identification of the impact of OBCs on fuel efficiency is simpler than that on truck age, namely because the timing issue is no longer as crucial. Whereas the effect of adoption on truck age takes time to become apparent, its seems reasonable to assume that the effect on mpg would manifest itself more quickly. Carriers can analyze information received from the OBCs, and then tell their drivers to adjust their behavior accordingly. This incentive effect on mpg is also most likely an 'intercept' effect, with no significant change once the initial benefit has been realized.

Table 8 provides the results for several different specifications that look at this relationship. The first specification that I consider is:

$$MPG_{it} = \lambda OBC_{it} + W_{it}'\zeta + \xi_{it} \tag{13}$$

where $MPG$ is average miles per gallon for a given truck; $OBC$ is one if an OBC is installed and zero otherwise; $W$ is a vector of control variables that influence mpg, including trailer

type, length of haul, principal product hauled, and truck age; $\xi_{it}$ is a white-noise error; and $\lambda$ and $\zeta$ are parameters. Though this specification is not dynamic, I again create a synthetic panel by aggregating up to the cell-level in exactly the same manner as in the truck age analysis. Studying this equation in cell-level form allows the results to be comparable to first-differenced specifications discussed below. The cross-sectional regressions for 1992 and 1997 are displayed in columns (1) and (2), respectively, of Table 8. Both estimates are positive and non-trivial, with a point estimate of .093 in 1992 and .112 in 1997, though neither is statistically significant at conventional levels. Columns (3) and (4) consider a first-differenced version of this specification:

$$\Delta MPG_{it} = \lambda \Delta OBC_{it} + \Delta W_{it}'\zeta + \Delta \xi_{it} \tag{14}$$

Here, the change in mpg is regressed on the change in OBC use, where the adoption of OBCs occurs somewhere in between the two points in time used to compute the difference in mpg. For example, column (3) looks at the changes in mpg between 1987 and 1992, and the trucks with OBCs in 1992 will have adopted at a point prior to 1992 (or they will have adopted in 1992 prior to taking the survey). If there is a one-time, immediate effect of OBCs on gas mileage, then the coefficient will capture the causal impact of the technology. Column (3) indicates basically a zero effect of OBC adoption on mpg, while column (4), which looks at the changes in variables between 1992 and 1997, estimates a positive and significant effect. Comparing columns (1) through (4) reveals an interesting pattern. The smallest coefficients are found in columns (1) and (3), which study OBC adoption in 1992, a date early in the diffusion of the technology. The largest estimates are in columns (2) and (4) where adoption is measured at a later date, in 1997, at which time the technology was well-established. These results are consistent with carriers needing time to learn how to use the technology, as they devise adequate measures of driver performance and implement

incentives to better shape behavior.

The regression in column (5) is a further test of the hypothesized intercept effect of OBC use on mpg. This regression is in the spirit of the truck age analysis, in which the growth in mpg in the future is regressed on the lagged adoption of OBCs. Not surprisingly, in this context the coefficient is imprecisely estimated, likely representing pure noise. The complete results of Table 8, particularly for the later years of OBC adoption, do indicate a positive effect of OBCs on fuel efficiency. Given that the average fuel efficiency of a truck is five mpg, the coefficient in column (4) of .149 suggests nearly a 3% improvement upon adoption of an OBC.[23]

# 7 Conclusion

This paper uses the introduction of a sophisticated monitoring technology in the trucking industry to test several implications of a principal-agent model. Two complementary empirical strategies are employed, one studying the variation in observed contracts and the other estimating the responsiveness of agent behavior to changes in monitoring. Suggestive evidence is provided that monitoring employees more intensively and providing a tighter linkage between performance and pay are complementary instruments at a firm's disposal. Stronger evidence is presented that employees' response to more precise performance measurement is quite elastic. The results imply that adoption of an OBC leads to substantial improvements in both truck life expectancy and fuel efficiency.

It should be noted that changes in the monitoring environment, and corresponding changes in incentive contracts, can have broader effects than simply on the efforts of given

[23]Baker and Hubbard's (2000) paper on asset ownership has a brief discussion of the effect of OBCs on mpg. They find that in a cross-section of data that OBCs increase fuel efficiency for company drivers more than for owner-operators, a result consistent with improved incentives.

drivers within firms. Longer-term mechanisms may be present, whereby increased monitoring of workers leads to a 'sorting' or 'selection' effect in which higher quality drivers are attracted to the industry and poorer quality individuals are terminated. This is certainly a possibility, and is consistent with the empirical work by Lazear (2000b) who finds significant selection effects under piece rates for windshield installers. The data used in this paper does not allow for separate identification of the incentive and selection components, so it is important to keep in mind that the improvements in truck life expectancy and fuel efficiency under OBC adoption are determined by these incentive and selection effects jointly. Future work should focus on the collection of appropriate data that would allow for the total change in the relevant outcome variables to be factored into these underlying structural components.

## References


[1] Aggarwal, Rajesh K., and Andrew A. Samwick. (1999). 'The Other Side of the Trade-off: The Impact of Risk on Executive Compensation.' *Journal of Political Economy*, 107 (1), pp. 65-105.

[2] Alchian, Armen A., and Harold Demsetz. (1972). 'Production, Information Costs, and Economic Organization.' *American Economic Review*, 62 (5), pp. 777-795.

[3] Asch, Beth J., (1990). 'Do Incentives Matter? The Case of Navy Recruiters.' *Industrial and Labor Relations Review*, 43 (February Special Issue), pp. 89S-106S.

[4] Baker, George, Michael Gibbs, and Bengt Holmstrom. (1994a). 'The Internal Economics of the Firm: Evidence from Personnel Data.' *Quarterly Journal of Economics*, 109 (4), pp. 881-919.

[5] Baker, George, Michael Gibbs, and Bengt Holmstrom. (1994b). 'The Wage Policy of a Firm.' *Quarterly Journal of Economics*, 109 (4), pp. 921-955.

[6] Baker, George P., and Thomas N. Hubbard, (2000). 'Contractibility and Asset Ownership: On-Board Computers and Governance in U.S. Trucking.' *NBER Working Paper 7634*.

[7] Baker, George P., and Thomas N. Hubbard (2001). 'Make Versus Buy in Trucking: Asset Ownership, Job Design and Information,' mimeo, University of Chicago Graduate School of Business.

[8] Baker, George P., Michael C. Jensen, and Kevin J. Murphy, (1988). 'Compensation and Incentives: Practice vs. Theory,' *Journal of Finance*, 43 (3), pp. 593-616.

[9] Belman, Dale A., and Kristen A. Monaco, (2001), 'The Effects of Deregulation, Deunionization, Technology, and Human Capital on the Work and Work Lives of Truck Drivers,' *Industrial and Labor Relations Review*, 54 (2A), pp. 502-524.

[10] Brickley, James A., and Jerold L. Zimmerman, (2002), 'Changing Incentives in a Multitask Environment: Evidence from a Top-Tier Business School,' *Journal of Corporate Finance*, (forthcoming).

[11] Cacciola, Stephen E., (2002), 'The Impact of a Monitoring Technology on Worker Incentives and the Coordination of Firm Activity: Evidence from the Trucking Industry', Chapter 1, Ph.D. Dissertation, Yale University.

[12] Chiappori, P.A., and B. Salanié, (2000), 'Testing Contract Theory: A Survey of Some Recent Work,' mimeo, University of Chicago.

[13] Deaton, Angus, (1985), 'Panel Data from a Time Series of Cross-Sections,' *Journal of Econometrics*, 30, pp. 109-126.

[14] Fernie, Sue, and David Metcalf, (1999), 'It's Not What You Pay it's the Way that You Pay and that's What Gets Results: Jockey's Pay and Performance,' *Labour*, 13 (2), pp. 385-411.

[15] Garen, John, (1994), 'Executive Compensation and Principal-Agent Theory,' *Journal of Political Economy*, 102 (6), pp. 1175-1199.

[16] Gaynor, Martin, and Paul Gertler, (1995), 'Moral Hazard and Risk Spreading in Partnerships,' *Rand Journal of Economics*, 26 (4), pp. 591-613.

[17] Gibbons, Robert, (1997), 'Incentives and Careers in Organizations,' in *Advances in Economics and Econometrics: Theory and Applications, Vol. 2*, ed. by David M. Kreps and Kenneth F. Wallis, Cambridge University Press.

[18] Hall, Brian J., and Jeffrey B. Liebman, (1998), 'Are CEOs Really Paid Like Bureaucrats?' *Quarterly Journal of Economics*, 113 (3), pp. 653-691.

[19] Haubrich, Joseph G., (1994), 'Risk Aversion, Performance Pay, and the Principal-Agent Model.' *Journal of Political Economy*, 102 (2), pp. 258-276.

[20] Higgs, Robert, (1973), 'Race, Tenure, and Resource Allocation in Southern Agriculture, 1910,' *Journal of Economic History*, 33 (1), pp. 149-169.

[21] Holmstrom, Bengt, (1979), 'Moral Hazard and Observability,' *Bell Journal of Economics*, 10 (1), pp. 74-91.

[22] Holmstrom, Bengt, (1982), 'Moral Hazard in Teams,' *Bell Journal of Economics*, 13 (2), pp. 324-340.

[23] Holmstrom, Bengt, and Paul Milgrom, (1987), 'Aggregation and Linearity in the Provision of Intertemporal Incentives,' *Econometrica*, 55 (2), pp. 303-328.

[24] Holmstrom, Bengt, and Paul Milgrom, (1991), 'Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design,' *Journal of Law, Economics, & Organization*, 7 (Special Issue), pp. 24-52.

[25] Holmstrom, Bengt, and Paul Milgrom, (1994), 'The Firm as an Incentive System,' *American Economic Review*, 84 (4), pp. 972-991.

[26] Hubbard, Thomas N., (2000), 'The Demand for Monitoring Technologies: The Case of Trucking,' *Quarterly Journal of Economics*, 115 (2), pp. 533-560.

[27] Hubbard Thomas N., (2001), 'Information Decisions and Productivity: On-Board Computers and Capacity Utilization in Trucking,' mimeo, University of Chicago Graduate School of Business.

[28] Jensen, Michael C., and Kevin J. Murphy, (1990). 'Performance Pay and Top-Management Incentives,' *Journal of Political Economy*, 98 (2), pp. 225-264.

[29] Lafontaine, Francine, and Scott E. Masten, (2002). 'Contracting in the Absence of Specific Investments and Moral Hazard: Understanding Carrier-Driver Relations in U.S. Trucking,' *NBER Working Paper 8859*.

[30] Lazear, Edward P., (1995), *Personnel Economics*, MIT Press.

[31] Lazear, Edward P., (1999a), 'Personnel Economics: Past Lessons and Future Directions,' *Journal of Labor Economics*, 17 (2), pp. 199-236.

[32] Lazear, Edward P., (1999b), 'Output-Based Pay: Incentives or Sorting?,' *NBER Working Paper 7419*.

[33] Lazear, Edward P., (2000a), 'The Future of Personnel Economics,' *The Economic Journal*, 110 (November), pp. F611-F639.

[34] Lazear, Edward P., (2000b), 'Performance Pay and Productivity,' *American Economic Review*, 90 (5), pp. 1346-1361.

[35] Medoff, James L., and Katherine G. Abraham, (1981), 'Experience, Earnings, and Performance,' *Quarterly Journal of Economics*, 95 (4), pp. 703-36.

[36] Mirrlees, James A., (1971), 'An Exploration in the Theory of Optimum Taxation,' *Review of Economic Studies*, 38, pp. 175-208.

[37] Mirrlees, James A., (1974), 'Notes on Welfare Economics, Information and Uncertainty,' in *Essays on Economic Behavior Under Uncertainty*, ed. by M. Balch, D. McFadden, and S. Wu, North-Holland Press.

[38] Nickerson, Jack A., and Brian S. Silverman, (1999), 'Why Aren't All Truck Drivers Owner-Operators? Asset Ownership and the Employment Relation in Interstate For-Hire Trucking,' *HBS Working Paper 00-015.*

[39] Ouellet, Lawrence J., (1994), *Pedal to the Metal: The Work Lives of Truckers*, Temple University Press.

[40] Prendergast, Canice, (1996), 'What Happens Within Firms? A Survey of Empirical Evidence on Compensation Policies,' *NBER Working Paper 5802.*

[41] Prendergast, Canice, (1999), 'The Provision of Incentives in Firms,' *Journal of Economic Literature*, 37 (1), pp. 7-63.

[42] Ross, Steven, (1973), 'The Economic Theory of Agency: The Principal's Problem,' *American Economic Review*, 63 (2), pp. 134-139.

[43] Shearer, Bruce, (2000), 'Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment,' mimeo, Université Laval.

[44] Spence, Michael, and Richard Zeckhauser, (1971), 'Insurance, Information, and Individual Action,' *American Economic Review*, 61 (3), pp. 380-387.

**Table 1**
**On-Board Computer Use Rates (in Percent) by Length of Haul and Trailer Type**

| | 1992 survey | 1997 survey |
|---|---|---|
| **Length of Haul:** | | |
| Off-Road | 3.05 | 10.72 |
| Local (50 miles or less) | 6.66 | 13.18 |
| Short Range (51 to 100 miles) | 11.27 | 20.25 |
| Short Range-Medium (101 to 200 miles) | 18.95 | 30.54 |
| Long Range-Medium (201 to 500 miles) | 24.36 | 39.73 |
| Long Range (501 miles) | 28.94 | 51.68 |
| **Trailer Type:** | | |
| Tank Truck | 21.72 | 36.76 |
| Refrigerated Van | 34.34 | 51.45 |
| Dry Cargo Van | 22.05 | 39.93 |
| Platform | 10.03 | 20.42 |
| Dump Truck | 7.34 | 16.15 |
| Grain Bodies | 5.41 | 10.45 |
| Other | 12.16 | 20.74 |
| All Trucks | 17.98 | 32.54 |
| (Number of Obs) | (39,850) | (25,533) |

Note: Statistics are based on truck-level data. Expansion factors provided by the Census are used as weights.

**Table 2**
**Summary of Bonuses and Base Pay Rates**

| | Number of Firms Offering | Mean Bonus Payment | Mean Base Pay Rate |
|---|---|---|---|
| **Productivity Bonus** | | | |
| None | 190 | ----- | $.310 per mile |
| Mileage | 46 | $.025 per mile | $.306 per mile |
| Lump Sum | 3 | $1600 per year | $.273 per mile |
| Percentage | 2 | 6% of haul revenue | 28% of haul revenue |
| **Performance Bonus** | | | |
| None | 172 | ----- | $.308 per mile |
| Mileage | 51 | $.019 per mile | $.305 per mile |
| Lump Sum | 15 | $1501 per year | $.315 per mile |
| Percentage | 3 | 2.3% of haul revenue | 25% of haul revenue |
| **Fuel Efficiency Bonus** | | | |
| None | 192 | ----- | $.311 per mile |
| Mileage | 40 | $.023 per mile | $.297 per mile |
| Lump Sum | 9 | $1707 per year | $.309 per mile |
| Percentage | 0 | ----- | ----- |
| **Safety Bonus** | | | |
| None | 125 | ----- | $.312 per mile |
| Mileage | 80 | $.016 per mile | $.304 per mile |
| Lump Sum | 29 | $730 per year | $.305 per mile |
| Percentage | 7 | 2.5% of haul revenue | 24.7% of haul revenue |

Notes: Statistics are computed from a firm-level data set of 241 for-hire carriers. Almost all of the firms offering bonuses paid in lump sum form pay base rates by the mile, so the per mile metric is used in the mean base pay rate column.

**Table 3**
**Correlation Coefficients of Driver Bonuses and On-Board Computer Use**

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| (1) Productivity | 1.000 | | | | |
| (2) Performance | .130 | 1.000 | | | |
| (3) Fuel Efficiency | .203 | .214 | 1.000 | | |
| (4) Safety | .049 | -.141 | .059 | 1.000 | |
| (5) OBC | .051 | -.002 | .222 | .111 | 1.000 |

Notes: Estimates are computed from cell-level data (112 cells) where the cell groupings are defined by length of haul, cargo (van) type, and base state.

**Table 4**
**OLS Regressions of Bonus Type on On-Board Computer Use**

| | Productivity | Performance | Fuel Efficiency | Safety |
|---|---|---|---|---|
| Unweighted Estimates: | | | | |
| OBC | .062 | -.020 | .276 | .203 |
| | (.147) | (.176) | (.144) | (.192) |
| Weighted Estimates: | | | | |
| OBC | .029 | -.080 | .177 | .241 |
| | (.135) | (.158) | (.148) | (.184) |
| Number of Cells | 112 | 112 | 112 | 112 |
| Number of Firms/ Trucks per Cell | 2.1/31.6 | 2.1/31.6 | 2.1/31.6 | 2.1/31.6 |

Notes: Standard errors are in parentheses. Each column represents a separate regression. Cells are constructed by length of haul, cargo (van) type, and base state. Covariates in all regressions include dummies for length of haul, cargo type, and base state. The weighted regressions are weighted by the sum of the number of firms (from the compensation data) and the number of trucks (from the OBC data) in a given cell.

**Table 5**
**Linear Probability Model Estimates of OBC Adoption as a Function of Truck Age and Other Operating Characteristics**

| | 1992 Survey | | 1997 Survey | |
|---|---|---|---|---|
| | Estimate | Stand. Error | Estimate | Stand. Error |
| 1 Year Old | **-.111** | **.017** | **-.034** | **.017** |
| 2 Years Old | **-.180** | **.016** | **-.079** | **.017** |
| 3 Years Old | **-.229** | **.015** | **-.152** | **.019** |
| 4 Years Old | **-.242** | **.015** | **-.169** | **.021** |
| 5 Years Old | **-.281** | **.015** | **-.219** | **.023** |
| 6 Years Old | **-.263** | **.016** | **-.282** | **.023** |
| 7 Years Old | **-.281** | **.015** | **-.375** | **.020** |
| 8 Years Old | **-.288** | **.015** | **-.379** | **.019** |
| 9 Years Old | **-.259** | **.019** | **-.403** | **.019** |
| 10 Years and Older (Omitted is New) | **-.283** | **.014** | **-.426** | **.016** |
| 50-100 Miles | **.019** | **.007** | .014 | .010 |
| 100-200 Miles | **.070** | **.009** | **.068** | **.013** |
| 200-500 Miles | **.104** | **.009** | **.089** | **.013** |
| Over 500 Miles (Omitted is <50 Miles) | **.096** | **.010** | **.134** | **.014** |
| Tank Truck | .023 | .015 | -.008 | .022 |
| Refrigerated Van | **.086** | **.013** | **.068** | **.017** |
| Platform Trailer | -.013 | .009 | -.024 | .014 |
| Specialized Trailer (Omitted is Dry Van) | -.004 | .009 | -.024 | .014 |

| | | | | |
|---|---|---|---|---|
| Truckload (Omitted is LTL) | **.131** | **.011** | **.130** | **.015** |
| Private Fleet | **.103** | **.012** | **.098** | **.016** |
| Contract Carriage (Omitted is Common) | **-.018** | **.009** | **.035** | **.011** |
| Fleet Size 25-99 | **.025** | **.007** | .017 | .011 |
| Fleet Size 100-499 | **.107** | **.010** | **.093** | **.013** |
| Fleet Size 500-999 | **.156** | **.017** | **.138** | **.021** |
| Fleet Size 1000-4999 | **.184** | **.014** | **.214** | **.019** |
| Fleet Size 5000-10000 | **.243** | **.022** | **.060** | **.025** |
| Fleet Size over 10000 (Omitted is <25 trucks) | **.132** | **.016** | **.142** | **.018** |
| Number of Trucks | 35,026 | | 22,122 | |

Notes: Expansion factors provided by the Census are used as weights. Estimates in bold indicate significance at the 5% level. Other covariates include base state, principal product hauled, intrastate operation, owner-operator, private refuel facility, and exempt carrier.

**Table 6**
**Measuring the Effect of On-Board Computer Use on Truck Age: OLS Relationships**

| Dependent Variable: | Age 1992 | Age 1997 |
|---|---|---|
| On-Board Computer 1992 | **-1.33**<br>**(.212)** | **-.524**<br>**(.227)** |
| Number of Cells | 4,299 | 4,299 |
| Mean number of trucks per cell | 7.82 | 6.50 |

Notes: Each column represents a separate regression. Standard errors are in parentheses. Cells are constructed by base state, principal product hauled, trailer type, and length of haul. Expansion factors provided by the Census and cell sizes are used as weights. Other covariates include base state, trailer type, principal product hauled, length of haul, truckload, private carriage, contract carriage, exempt carrier, owner-operator, intrastate operation, private refuel facility, and fleet size.

**Table 7**
**Measuring the Effect of On-Board Computer Use on Truck Age: Corrections for Endogeneity**

| | First Difference | Instrumental Variables |
|---|---|---|
| **Dependent Variable:** | **Age Growth 1992 to 1997** | **Age 1997** |
| On-Board Computer 1992 | **.952**<br>**(.271)** | **1.21**<br>**(.689)** |
| Number of Cells | 4,299 | 4,175 |
| Mean number of trucks per cell | 6.50 | 7.36 |

Notes: Each column represents a separate regression. Standard errors are in parentheses. Cells are constructed by base state, principal product hauled, trailer type, and length of haul. Expansion factors provided by the Census and cell sizes are used as weights. The excluded instruments used in the IV strategy include the four accident variables available in the 1987 survey, as well as these accident variables interacted with the product type dummies. This results in 73 excluded instruments. Other covariates include base state, trailer type, principal product hauled, length of haul, truckload, private carriage, contract carriage, exempt carrier, owner-operator, intrastate operation, private refuel facility, and fleet size.

**Table 8**
**Measuring the Effect of On-Board Computer Use on Miles Per Gallon**

| Dependent Variable: | MPG 1992 | MPG 1997 | MPG Growth 1987 to 1992 | MPG Growth 1992 to 1997 | MPG Growth 1992 to 1997 |
|---|---|---|---|---|---|
| Covariates: | | | | | |
| OBC 1992 | .093<br>(.076) | --- | -.037<br>(.124) | --- | -.155<br>(.128) |
| OBC 1997 | --- | .112<br>(.070) | --- | --- | --- |
| OBC Growth 1992 to 1997 | --- | --- | --- | .149<br>(.067) | --- |
| Number of cells | 4,200 | 4,086 | 4,150 | 4,200 | 4,200 |
| Mean # of trucks per cell | 7.94 | 5.39 | 6.62 | 6.61 | 6.61 |

Notes: Each column represents a separate regression. Standard errors are in parentheses. Cells are constructed by base state, principal product hauled, trailer type, and length of haul. Expansion factors provided by the Census and cell sizes are used as weights. Other covariates include base state, trailer type, principal product hauled, length of haul, truckload, private carriage, contract carriage, exempt carrier, owner-operator, intrastate operation, private refuel facility, fleet size, and truck age.

**Appendix Table 1**
**First Stage Results of Instrumental Variables Strategy**
**Dependent Variable: OBC Use in 1992**

| | Estimate | Standard Error |
|---|---|---|
| Number of Accidents in 1987 (Acc) | **-.535** | **.187** |
| Number of Fatalities in 1987 (Fatal) | **1.428** | **.208** |
| Number of Bodily Injuries in 1987 (Inj) | .961 | .716 |
| Number of Property Damages >$4,200 in 1987 (Prop) | **.378** | **.190** |
| | | |
| Farm Products | **-.074** | **.028** |
| Live Animals | **-.078** | **.032** |
| Processed Food | .017 | .032 |
| Mining Products | .023 | .050 |
| Building Materials | -.036 | .028 |
| Logs and Forest Products | .007 | .031 |
| Lumber and Fabricated Wood Products | -.040 | .033 |
| Paper Products | .024 | .046 |
| Petroleum | -.008 | .029 |
| Plastics and Rubber | -.046 | .047 |
| Primary Metal Products | **-.106** | **.042** |
| Fabricated Metal Products | -.053 | .033 |
| Machinery | **-.061** | **.028** |
| Transportation Equipment | -.011 | .038 |
| Furniture and Hardware | **-.134** | **.038** |
| Textile Mill Products | -.004 | .048 |
| Household Goods | .011 | .054 |
| Craftman's Equipment | -.080 | .094 |
| Mixed Cargo | -.061 | .044 |
| Other Cargo | -.020 | .028 |
| | | |
| Acc*Farm Products | **.682** | **.231** |
| Acc*Live Animals | **.576** | **.220** |
| Acc*Processed Food | **.540** | **.226** |
| Acc*Mining Products | **.428** | **.214** |
| Acc*Building Materials | **.530** | **.218** |
| Acc*Logs and Forest Products | **.516** | **.197** |
| Acc*Lumber and Fabricated Wood Products | **.873** | **.262** |
| Acc*Paper Products | .538 | .288 |
| Acc*Petroleum | **.782** | **.229** |
| Acc*Plastics and Rubber | **1.565** | **.560** |
| Acc*Primary Metal Products | **1.256** | **.287** |
| Acc*Fabricated Metal Products | .312 | .211 |
| Acc*Machinery | **.721** | **.230** |
| Acc*Transportation Equipment | **.777** | **.364** |
| Acc*Furniture and Hardware | **.690** | **.266** |
| Acc*Textile Mill Products | -.040 | .355 |

| | | |
|---|---|---|
| Acc*Household Goods | .433 | .250 |
| Acc*Craftman's Equipment | .151 | .105 |
| Acc*Mixed Cargo | **.632** | **.263** |
| Acc*Other Cargo | **.487** | **.206** |
| | | |
| Fatal*Farm Products | **-1.025** | **.377** |
| Fatal*Processed Food | -1.137 | .647 |
| Fatal*Building Materials | **-1.258** | **.653** |
| Fatal*Logs and Forest Products | **-1.508** | **.334** |
| Fatal*Lumber and Fabricated Wood Products | **-1.203** | **.552** |
| Fatal*Petroleum | **-1.925** | **.492** |
| Fatal*Plastics and Rubber | -1.748 | .965 |
| Fatal*Fabricated Metal Products | **-1.478** | **.250** |
| Fatal*Machinery | **-2.328** | **.269** |
| Fatal*Textile Mill Products | -.921 | .614 |
| Fatal*Mixed Cargo | **-1.777** | **.844** |
| Fatal*Other Cargo | **-2.175** | **.320** |
| | | |
| Inj*Farm Products | -1.000 | .733 |
| Inj*Live Animals | -.517 | .988 |
| Inj*Processed Food | -.386 | 1.042 |
| Inj*Mining Products | .287 | .823 |
| Inj*Building Materials | -1.029 | .735 |
| Inj*Logs and Forest Products | -1.234 | .733 |
| Inj*Lumber and Fabricated Wood Products | -1.137 | .770 |
| Inj*Paper Products | .617 | .916 |
| Inj*Petroleum | -1.185 | .736 |
| Inj*Primary Metal Products | -.765 | .868 |
| Inj*Fabricated Metal Products | .338 | .757 |
| Inj*Machinery | -1.260 | 1.183 |
| Inj*Transportation Equipment | **-6.140** | **2.577** |
| Inj*Furniture and Hardware | -2.833 | 2.465 |
| Inj*Textile Mill Products | .491 | 1.141 |
| Inj*Household Goods | -1.124 | .769 |
| Inj*Mixed Cargo | -.415 | .966 |
| Inj*Other Cargo | -.665 | .735 |
| | | |
| Prop*Farm Products | **-.530** | **.224** |
| Prop*Live Animals | -.318 | .304 |
| Prop*Processed Food | -.622 | .451 |
| Prop*Mining Products | -.357 | .324 |
| Prop*Building Materials | **-.635** | **.255** |
| Prop*Logs and Forest Products | -.355 | .221 |
| Prop*Lumber and Fabricated Wood Products | -.213 | .292 |
| Prop*Paper Products | -.801 | .493 |
| Prop*Petroleum | -.215 | .370 |
| Prop*Plastics and Rubber | -1.050 | .564 |
| Prop*Primary Metal Products | **-.895** | **.434** |
| Prop*Fabricated Metal Products | -.149 | .229 |

| | | |
|---|---|---|
| Prop*Machinery | -.695 | .363 |
| Prop*Transportation Equipment | **-1.097** | **.500** |
| Prop*Furniture and Hardware | -.924 | .781 |
| Prop*Textile Mill Products | .050 | .346 |
| Prop*Household Goods | -.366 | .247 |
| Prop*Mixed Cargo | -.790 | .408 |
| Prop*Other Cargo | **-.642** | **.221** |
| F-statistic for joint significance of the excluded instruments (p-value) | F(73, 4010) = 5.48 (0.0000) | |

Notes: The product type main effects are **not** excluded instruments; they are included in both the first and second stage regressions. The excluded instruments are the four accident variable main effects and the accident variable and product dummy interactions. Cells are constructed by base state, principal product hauled, trailer type, and length of haul. Expansion factors provided by the Census and cell sizes are used as weights. Estimates in bold indicate significance at the 5% level. The omitted product type category is chemicals. Other covariates include base state, trailer type, length of haul, truckload, private carriage, contract carriage, exempt carrier, owner-operator, intrastate operation, private refuel facility, and fleet size.

# Chapter 3
# Inside the 'Black Box' of Project STAR: Estimation of Peer Effects Using Experimental Data

**by Michael A. Boozer and Stephen E. Cacciola**

## 1 Introduction

The question of the existence and the quantitative importance of peer effects in influencing individual behavior has long eluded credible empirical study. The essential problem is that whether the researcher is interested in how individual behavior is affected by group characteristics (termed exogenous or contextual effects) or group behavior (termed endogenous effects), data are rarely available in which the relevant groups or their associated traits are exogenously assigned. While this criticism applies to *any* empirical study when we examine how individual traits are associated with individual outcomes, the problem is particularly vexing in the study of peer effects. The conceptual problems are numerous, and well elucidated in the literature (see especially the writings of Manski (1993, 1995, 2000) in the economics literature, and Hauser (1970) in the sociology literature) and indicate the numerous pitfalls whereby a researcher may erroneously infer the presence of peer effects, when in fact the estimates may only be indicative of the respondent and her associated group sharing a common environment.

As the conceptual idea related to the study of peer effects places the same individual in a variety of alternative group settings (based either on (exogenous) inputs or outcomes, depending on what is of interest to the researcher), the ideal data required by the empirical researcher needs to sample a large number of nearly identical individuals placed in a multiplicity of alternative group settings. The problem is how to mimic this conceptual ideal

with observational data, whereby alternative group settings almost surely carry with them differences based on unobserved characteristics as well.

Here again, the problem of the unobservables confounding inference is clearly not unique to the study of peer effects. But as one of the canonical modes of detecting and quantifying the importance of peer effects places some measure of group outcomes as one of the key explanatory factors in a regression for individual behavior, the presence of these unobservables becomes particularly acute. In particular, even if we can argue that the other covariates in such a regression are plausibly exogenous, to the extent that the unobservables are shared by some or all of the other group outcomes, then the summary measure of the group outcomes that serves as the peer effect measure will appear spuriously important for that reason. Thus, the criteria that must be imposed on the unobservables in order for the researcher to claim that the estimated peer effects represent something of *behavioral* significance (as opposed to simply representing a quantified version of the statement that they all share a common environment) are far more stringent than for a simple regression which is used to understand individual attributes and individual outcomes.

We take up this challenge in this paper by utilizing data on an experiment conducted in Tennessee in the early 1980's designed ostensibly to study the effects of class size on student achievement in grades Kindergarten through third grade. These data are commonly called the Project STAR data, and they have been studied extensively in the literature with regards their to primary objective, the class size effect. Some of the more well-cited papers include Krueger (1999), Hanushek (1999), and Finn and Achilles (1990). We argue that the effects found by these authors represent a reduced-form impact of class size, but that they do not try to break these effects down into their constituent components. In particular, we take the view that Heckman (1992) has offered on social experiments generally, in that they constitute a 'black box' of underlying components. Heckman has pointed out that it

is essential to understand these more structural components of social experiments so as to properly extrapolate the knowledge gained from them to large-scale policy implementation. In our work here, we focus on the crucial aspect of Project STAR in that it was conducted over several grades. As the experiment progressed over time, from Kindergarten to third grade, it is possible that the experimental effects capture less a 'pure' class size effect and potentially more a feedback effect (or 'social multiplier'), operating through the experimentally induced peer quality differences across classes. It is important to note that we do not disagree with the authors who have written on the Project STAR results as regards the reduced form results they find and report, but we do offer an alternative interpretation of these results in such a way that allow for quite different policy proposals (i.e. not based *entirely* on changing class sizes) which may offer the same slate of academic outcomes.

At the core of our reinterpretation of the Project STAR results is the main purpose of this paper, which is to estimate peer effects using data wherein some fraction of the variation in reference group characteristics is exogenously determined. We are interested in this paper in 'endogenous' peer effects (as termed by Manski) whereby individual outcomes are altered by some aspect of the distribution of the reference group outcomes. Such peer group effects have the feature that they generate a feedback effect, so that the intensity to which social programs operate within and between groups affects the total aggregate outcome. Positive feedback, for example, would imply that social programs which are highly concentrated on groups of individuals will be more efficient than programs which are 'sprinkled' across the landscape. While the Project STAR design in principle kept students assigned to Small classes in Small classes for the duration of the experiment (and the same for the students in Regular sized classes), the exit and subsequent replacement of students from and into the Project STAR schools meant that the population of students participating in the experiment had differential exposures to the Small and Regular class size treatments. This fact is the

key to our identification strategy for the estimation of the peer groups effects.

The simultaneous determination of an individual student outcome and her corresponding class group outcomes, as well as their common exposure to a class size of a given type (Small or Regular), both necessitate that we need a means by which we can use the experimental design to deliver an instrumental variable(s) by which some fraction of the variance in group outcomes is exogenously determined. Were students exogenously assigned to not just class *types* within schools, and were test scores available for the newly entering students *before* they enrolled in the Project STAR schools, we could simply utilize ordinary least squares, using a measure such as the sample mean of the *lagged* test scores of a student's current classmates as the peer group measure. While this approach is not possible owing to the lack of test scores for the new entrants, this idea does emphasize the value of the longitudinal nature of the experiment. In particular, a suitable version of *previous* exposure to the Small class treatment is a good candidate for an instrument. At the individual level, this prior exposure to the treatment is a component of lagged test scores that we can observe, and so using the fraction of the class previously exposed to the Small class treatment is a suitable candidate instrument for the student's current peer group average test scores. The fact that the instrument is lagged is what allows us to avoid the simultaneous determination of the individual student's outcome, as well as the outcomes of her peers. This idea utilizes the experimental design to extract the variation in student performance due to the impact of the experiment in an earlier grade, because of the boost in performance owing to the Small class treatment versus both the Regular class treatment and the entire group of newly entering students who had no prior exposure to the experiment.

This is where the exit, and subsequent replenishment, of students out of and into the Project STAR schools is crucial for our purposes. In the extreme case where no exit and entry takes place, then our instrument for peer group quality would be perfectly collinear

with the class type indicator, and we would be unable to infer what is a peer group effect from what is a class type effect.[1] Fortunately, the entry and exit patterns of students across classes as well as across schools was quite diverse, and so we have rather good power in explaining group outcomes, even conditional on a class type indicator included as a regressor. We interpret the coefficient on the class type regressor as a 'pure' class type (or size) effect, net of the feedback effects due to alterations in peer group quality from the impact of the experiment in the earlier grades. Not surprisingly, owing to the lag nature of our strategy to split these two effects apart given the overall reduced form effect, we have no power to tell these apart for Kindergarten, and extremely little power to do so as of the first grade. However, for the second and third grades, we have relatively good power, and we find that after controlling for the experimentally determined peer group effect, the pure class size effect is rendered much smaller than the reduced form effects found in the earlier studies on Project STAR, and in many cases, these 'pure' class size effects are insignificantly different from zero. The bulk of the reduced form effects as of the second and third grades appears to be due to the feedback of the peer group effects.

We also comment on the methods used to estimate the importance of peer group effects commonly used in the literature, and link these to methods used to study phenomena which may be quite distinct from the study of peer group effects. Fundamentally, peer group effects are spillover effects whereby group output exceeds individual effects summed to the group level. The degree to which the per-person group output exceeds the individual output is the peer effect. We show that this is precisely what is estimated by the canonical

[1] In fact, this is also a version of the 'reflection problem' (as labeled by Manski (1993)) whereby it is unclear what fraction of students performing well in a Small class is due to the class size effect as opposed to the peer group effect. Absent entry and exit of students from the Project STAR schools, we would be unable to apportion what fraction of a class type effect is due to a pure resource effect, and what fraction is due to a peer effect.

approach in the literature which estimates variants of regressions of individual outcomes on typically the average of the outcomes of the other members of the peer group. We also discuss the specification problems which lead to meaningless coefficients of 1 in extreme circumstances, but possibly less than 1 (but with no more meaning) in more typical settings, thereby obscuring the spurious regression problems plaguing the research exercise. We then consider a variety of alternative means by which peer group effects may be estimated from the data, as well as specification checks that can be performed.

The next section of the paper discusses the Project STAR experimental design and the aspects of the data which are crucial for our research question. We then provide a brief conceptual discussion in Section three of the identification issues involved in extracting the peer group effects from the Project STAR data. In Section four we discuss our core empirical results. Section five then considers the more conceptual issues involved in the estimation of peer effects generally, and Section six concludes.

## 2 The Project STAR Experimental Design and Data

Project STAR was funded by the Tennessee State Legislature and conducted by the Tennessee Department of Education with the goal of obtaining conclusive results regarding the efficacy of class size reductions.[2] The ambiguity of the existing empirical literature, which used observational data, compelled the Legislature to appropriate funding in order to design, implement, and interpret an experimental study before investing in across-the-board slashing of class sizes. The 79 schools that participated in the first year of the study, the 1985-86 school year, were selected to provide variation in both geographic location across the state and in the size and economic status of the school locations (schools were designated

[2]For more comprehensive descriptions of the experiment see Folger (1989), Word *et al.* (1990), Finn and Achilles (1990), and Krueger (1999).

as inner city, suburban, urban, or rural). Importantly, the experimental randomization took place within schools, so that participating schools were required to be large enough to have at least one class of each type under study. At the outset of the experiment, kindergarten students and their teachers were randomly assigned to one of three class types: Small classes (13-17 students), Regular classes (22-25 students), or Regular/aide classes (22-25 students) which included a full-time teacher's aide.[3] The experimental design called for students to remain in the same class type through the end of third grade, at which time all children returned to Regular size classes. Students entering STAR schools after kindergarten were added to the experiment. All told, there were between 6,000 and 7,000 students in the experiment in each year, and the experiment involved a total of 11,600 children over all four years.

The validity of any experimental study may be compromised if the random assignment is not credible. As such, the schools participating in the STAR experiment were audited to enforce compliance with the random assignment procedures. A critical piece of our identification of peer group effects lies with the new students who entered the participating schools during the course of the STAR experiment. Fortunately, the protocol was for all entering children to be randomly assigned to a class type. All available indications are that the initial random assignment to classes of students, both those attending kindergarten as well as those entering in later grades, and teachers was done soundly. Since the STAR data only contains information on the actual class type a student attended in a given year, and not the type of class to which the student was randomly assigned, Krueger (1999) explores the possibility that students switched class types immediately after their random

[3]The average class size over the course of the experiment was 15.3 for the Small classes, 22.8 for the Regular classes, and 23.2 for the Regular/aide classes. In the 1985-86 school year, the statewide pupil-teacher ratio in Tennessee was 22.3.

assignment. In his subsample of 1581 students in 18 schools, he finds that for 99.7% of students, the class type attended in kindergarten was the class type to which the students were randomly assigned. This indicates that the initial random assignment of students was taken very seriously by the participating schools.

Note also that if the randomization were done correctly, we would expect the average characteristics of students across the treatment and control groups to look identical prior to the start of the experiment. Unfortunately, students were not given a baseline test before attending class, so it's not possible to compare test scores across class type to address credible randomization. But we can of course compare the observable characteristics of students (as well as teachers) and see if on average they look similar in Small, Regular, and Regular/aide classes. Krueger and Whitmore (2001) performed this exercise for both students and teachers. For students, class-type assignment was modeled as a function of demographic characteristics (free lunch[4], race, and gender) and school-by-entry-wave fixed effects to account for the fact that randomization occurred within schools and at the time in which a student entered the experiment. The results indicate that student characteristics are not correlated with assignment status, as we would expect under random assignment to class type. An analogous model was estimated for the assignment of teachers, with the relevant demographic characteristics being race, gender, master's degree, and total experience. Again, these characteristics are not jointly significant in explaining assignment status, a result consistent with the random placement of teachers in class types.

As is common in social experiments, particularly those of an extended longitudinal nature, Project STAR deviated both in its administration and due to behavioral responses of the participants in a way that was not ideal given the intentions of the original experimental design. Rather than weakening the merit of the experiment, we argue that in this case

[4]Free lunch is intended as a measure of parents' economic status.

particular exogenous changes in the composition of classes allow us to address a broader set of issues than solely the effectiveness of class size reductions. The first deviation, and of only limited interest in our analysis, is at the end of kindergarten students in Regular and Regular/aide classes were re-randomized between these two class types. In a practical sense, the distinction between Regular and Regular/aide classes is inconsequential since many of the Regular classes employed a part-time aide. Empirically, the results of the Project STAR experiment indicate that the difference in student achievement between Regular and Regular/aide classes is insignificant. Nonetheless, in our analysis that follows we often distinguish between Regular and Regular/aide classes when modeling student outcomes, but our principal instrument for peer quality groups Regular and Regular/aide students together.

A second departure from the original experimental protocol is that a number of students, on the order of 10% per year, switched between Small and Regular classes. Krueger (1999) attributes this primarily to behavioral problems and parental complaints. If the students who switched class types systematically differed from those who remained with their initial assignments, then a comparison of outcomes of the treatment and control groups may no longer estimate a parameter of interest.

Finally, student mobility substantially affected the experimental design. Students attrited out of the experiment, due in part to families having moved to different school districts and students having attended private schools, and students entered STAR schools after kindergarten. Since kindergarten was not mandatory in Tennessee at the time of the experiment, a particularly large influx of students is seen entering in first grade (2313 new students entered in first grade compared with 4516 of the kindergarten students remaining in the experiment at that time). A substantial number of new entrants also appeared later in the experiment; 1679 students entered in second grade and 1281 students entered

in third grade. We argue that it is primarily this inflow of new students that renders a simple comparison of treatment and control groups ineffective in isolating the 'pure' class size effect. To credibly estimate the class size effect, it is also necessary to consider the difference in peer group composition induced by the new entrants and, to a lesser extent, the students switching between class types. More specifically, the new entrants generate variation in peer quality via two distinct routes. First, a new entrant does not have the 'boost' in achievement provided by attendance in a Small class, so if the student is randomly assigned to a Small class he lowers the average quality of students in that class. Second, the STAR data indicates that students who entered the experiment after kindergarten are lower achievers than those who attended STAR schools at the outset of the experiment. This may occur because the late entrants did not attend kindergarten, and may also reflect unobserved family background characteristics and parents' tastes for their childrens' education. The new entrants are then randomly assigned to a class type, and 'water-down' the quality of both the Small and Regular classes.

Table 1 summarizes the mean characteristics of students in the sample by their transition status between grades[5]; students either switch class type, remain in the same class type, or are new entrants into the experiment. A comparison of the 'switchers' with the 'stayers' indicates that the movement of students between class types is likely nonrandom. Comparing the switchers to those who remain in their initially assigned class type, we see that the switchers tend to have a slightly higher tendency to be on free lunch. But the comparisons between gender and race reveal essentially no systematic differences. On average, students who switched from a Small class to a Regular class between grades had lower test scores prior to switching than those students remaining in a Small class. The averages in

[5]Net of the variation across schools. Because the schools themselves were not selected at random, all analyses in this paper condition on school effects.

Table 1 also illustrate the disparities between the group of new entrants and the students previously in the experiment. In addition to lower test score averages, new entrants are more likely to be nonwhite, male, and on free lunch than students already attending STAR schools.

Given the probable nonrandom selection of students who switch class type, we emphasize that we primarily identify the peer group effects off of the variation induced by the new entrants. Table 2 lists the number of students in each grade and class type by the students' place of origin: randomly assigned to the relevant class type, switched from the other class type, or new entrant. The number of students previously randomly assigned to their current class type dominate the switchers, consistent with the experimental protocol for students to remain in the same class type through the end of third grade. The new entrants substantially outnumber the switchers in any given year, lending credence to our identification strategy.

This study uses the Project STAR Public Access Data, which follows the initial cohort of participating students, plus new entrants, through third grade. The data contains student level observations and includes the whole universe of students in the experiment in a given year, not just a subsample. The key variables included for each observation are student characteristics (race, gender, free lunch status), teacher characteristics (race, hold master's degree, total experience), class type, school identifiers, and test scores. The Public Access Data contains two test scores: the Stanford Achievement Test (SAT) in reading and the SAT in math, which were administered to students at the end of each school year. Following Krueger (1999), we rescaled the raw test scores into percentiles. For each grade and test measure, the Regular and Regular/aide students were grouped together and given percentile scores ranging from 0 to 100. The students in Small classes were then assigned a percentile score for each test based on where their raw scores fell in the distribution of Regular-class students. To obtain the percentile test score measure used in our analysis, we took the

average of the percentile math score and the percentile reading score.[6] If one of these scores was missing, we used the one available score as the percentile test score.

Our analysis for estimating peer group effects requires knowing which students were taught in the same class. The Public Access Data only identifies class *type*, so if, for example, there was more than one Small class in a school, we had to infer which students were grouped together and physically located in the same classroom. We did this by using the teacher characteristics variables collected for each student. If students in the same school and class type had been taught by, say, a white teacher with a master's degree and 25 years of total experience, we could safely assume that these students were classmates.[7]

# 3 The Identification of Peer Group Effects With the Project STAR Data

Before moving to a more general discussion of issues and alternative methods of the estimation of peer effects, we begin with a simplified discussion of how we use the Project STAR data to estimate standard peer group effects. The canonical regression model that has been used in the literature to study peer group effects (of the typed coined 'endogenous' by Manski) is usually a variant of:

$$y_{ij} = \beta \bar{y}_{-i,j} + x'_{ij}\gamma + \epsilon_{ij} \tag{1}$$

[6]Krueger (1999) has access to several additional tests: the SAT word recognition test, and the Tennessee Basic Skills First (BSF) tests in reading and math. His primary analysis uses the SAT word recognition test in addition to the SAT reading and math tests. Our ability to replicate his results indicates that the absence of the SAT word recognition test in our data is of little consequence.

[7]In a few cases, it appears that two teachers in the same school and teaching in the same class type did have identical characteristics. For their students, we could not determine which ones were grouped together, so these students were dropped from our analysis in the relevant grade. This resulted in our losing 77 students in kindergarten and 47 students in the third grade.

where $y_{ij}$ is the outcome of interest for individual $i$ who has group affiliation $j$. As is typical in this literature, we start by assuming that the peer group affiliation is known *a priori* by the researcher, and in our case, we assume it is the student's classroom.[8] The key regressor of interest is the sample mean of the group outcomes, net of individual $i$'s outcome, a quantity commonly referred to as the 'leave-out mean' denoted as $\bar{y}_{-i,j}$ where

$$\bar{y}_{-i,j} \equiv \frac{1}{N-1}\sum_{k\neq i}^{N-1} y_{kj} = \frac{1}{N-1}(N\bar{y}_j - y_{ij}) \tag{2}$$

For ease of exposition, we have assumed that the group sizes are the same across groups and it is designated by $N$. Indeed, in the Project STAR data, within a class *type* subgrouping, the class size $N$ is ideally homogeneous, but in fact it does vary. We let $J$ denote the number of groups, and so the sample size in this simplified setup (ignoring the differences in class sizes) is $NJ$. Also, the fact that the data include *every* individual in a given class implies that we can use the leave-out mean as the peer group measure. In typical observational datasets such as the High School and Beyond, or the National Education Longitudinal Study (NELS), only a small fraction of a student's peers in a school are included in the survey, and so researchers would often use the group mean *inclusive* of individual $i$, $\bar{y}_{ij}$, as that was more representative of the population-level mean outcome for the school. The nature of the Project STAR data affords us the luxury that we do not have to deal with some of the issues that arise when using the group mean inclusive of individual $i$'s outcome when studying the determinants of $y_{ij}$.

[8] An extremely small minority of work on this topic tries to confront this issue seriously, as opposed to replacing our residual ignorance of peer group affiliation with blunt force assumptions needed to make the research venture progress. Woittez and Kapteyn (1998) use survey responses as to who constitutes peers as the relevant peer group, as opposed to simply assigning generic group designations as we have done. Conley and Udry (2000) use survey responses on conversations about farming methods to deal with learning models in development. Manski (2000) points out the formidable identification problems when group affiliation is not known *a priori*.

While the canonical approach has taken the mean of reference group behavior as the relevant peer group measure, here again this is done for lack of information as to what features of the distribution of peer group outcomes are relevant for individual behavior. It could be the 90th percentile, or the 10th percentile, or possibly not just the mean, but perhaps also lower variance aids in enhancing individual achievement *ceteris paribus*. We agree these are unsolved and interesting issues, but again ignore them for the moment, and focus on identification issues with the set of canonical assumptions.

The point is that even with the litany of strong assumptions we have already imposed, the problem of identifying $\beta$ from the above equation is still not nearly solved. The essential problems are two-fold: (i) The individuals who comprise each peer group $j$ are not generally exogenously (as regards individual outcomes) determined and (ii) even when groups are exogenously formed (by a lottery or some randomization device), individual and group outcomes are simultaneously formed, a problem termed the 'reflection problem' by Manski as an analogy to a mirror image thought to be *causing* its corresponding object to move, as opposed to be simply reflecting it. As we indicated above, the reflection problem implies that simply estimating equation (1) without regard to this issue implies nothing more than a quantitative statement that the individual and the peer group share a common environment.

To move beyond such statements and to try to capture the *behavioral* impacts of a peer group on individual behavior, we need an empirical strategy which will abstract from the two prominent sources of endogeneity just discussed. The question of peer group formation is a common issue in empirical economics as it is just a form of sorting or endogenous migration. Perhaps one of its best known forms is that of Tiebout sorting wherein the demand for public goods across communities needs to first address *why* those communities formed in the first place. The general strategy in such situations is to either try to find some fraction of the variance in group composition which is exogenously determined, or to

exploit variation in the public good demand which is not determined by the preferences of communities. Alternatively, one could try to fully model the process by which groups are formed, and thereby use sources of variation from that model which are unrelated to the outcome process. Unfortunately, this latter approach requires very rich data on preferences as well as detailed data on group members and potential group members, or it runs the risk of being a tautological exercise in that it faces little discipline from the data.

The flip-side of this concern over the endogeneity in the peer group measure $\bar{y}_{-i,j}$ is also ensuring that a suitable instrument is also correlated with the peer group measure, *net* of the other covariates. This is the rank condition necessary for identification, and the key issue here is that it has to hold in the presence of the covariates. This is not trivial, as one of the key regressors is the indicator for whether the child is assigned to a Small class in her current grade, which we label $D_j$. We let $D_j = 1$ when the child is assigned to the Small class treatment, and clearly, for a given class $j$, this does not vary at the individual student level.[9] Therefore, any peer group measure, or any candidate instrument for the peer group measure, must vary within classes in order to satisfy the rank condition. Naturally, this would rule out, for example, differences in peer group measures *between* the treatment and control groupings of the Small and Regular classes. The problem with such an identification strategy is that we would be unable to distinguish between what is a pure class size effect versus what is a peer group effect as the two measures move completely in tandem within schools.

In order to drive a wedge between the current class-size designation category $D_j$ and some factor which uses the experiment to generate exogenous changes in peer group com-

[9]The reader should also bear in mind the experiment did not utilize a random selection of schools, as discussed in Section 2 above. As such, all econometric methods implicitly contain a set of school fixed-effects. For that reason, only instruments that contain some within-school variation are valid candidates to use as instrumental variables.

position, we turn instead to the *timing* of the experimental impacts and the essence of the feedback effect. As we discussed in Section 2, the exit of children from the Project STAR schools and the subsequent random assignment of children to Small and Regular classes to fill their place imply that a child who is randomly assigned to a Small class in her current grade was not necessarily in the Small class in the previous year if she was new to the Project STAR schools. In order to avoid cluttering the notation with an additional subscript denoting the timing of variables, let us stick to our current notation scheme (of labeling things for the current grade only), but define a new variable for the children of class $j$ to indicate their random assignment status for the *previous* class year $d_{ij}$. Therefore, $d_{ij} = 1$ if student $i$ was *previously* randomly assigned to a Small class, and due to the exit and entry of students, it is not necessarily the case that in Small classes (i.e. $D_j = 1$) that $d_{ij}$ is 1 for each student.[10] As a useful piece of additional notation, define the number of students in each class $j$ who were previously randomly assigned to a Small class as $S_j \equiv \sum_{i=1}^{N} d_{ij}$, and the associated fraction of students who were previously randomly assigned to a Small class $z_j \equiv \frac{1}{N} S_j$.

Now if *all* students in the current class $j$ had valid test score measures taken before they began the year in class $j$, then we could use this average as one measure of the peer group quality and study the impact of this measure on individual test scores at the end of the school year. However, even apart from the fact that we only have such data for *incumbent* participants in the Project STAR study, this simple but direct approach would have potential pitfalls. First, while it is true that students were randomly assigned to class *types*, it is not clear they were randomly assigned to specific classes within the class type

[10] We are ignoring the rather small fraction of students who switch class type assignments in violation of the experimental protocol. They are not essential to our identification strategy, and they only add inessential complexity to incorporate them into our current discussion.

categories within schools. Second, the OLS approach of using the lagged average of test scores on the student's current year peers assumes the other inputs to the test score outcome that are common to the entire group are controlled for in the regressors. In fact, even with the measure under study, class size, there were small but detectable differences in class sizes within a given class type category. Thus, even with the use of the lagged measure, we may have to be careful to avoid an omitted variables problem when looking across years. Finally, we come back to the reality of the data that we lack test scores for the previous year for the New Entrants, and so they would have to dropped in order for such an analysis to be feasible.

Instead, we make use of the hypothesis that the class size treatment *assignment*[11] had an impact on the subsequent year's test score to solve these three problems. In particular, by grouping the New Entrants with the Regular class students and contrasting them with the 'boost' in test scores received by the children placed in Small classes in the previous year, we can conceptually extract the component of the lagged test score that was induced by the experiment by using the variation in *current* scores explained by lagged treatment status. Furthermore, as regards the possible failure of the exogenous assignment of students to individual classes within class types, we can replace this with the somewhat weaker assumption that the class groupings were not determined by the fraction of children previously randomly assigned to Small classes. Finally, as regards the possible omitted variables common both to the student and her peer group, now we need to only worry about omitted variables that are correlated with the fraction of children in each class that were previously assigned to the Small class types. Of course, as we do not have any explicit randomization device creating the classes, we cannot be positive some type of exogeneity failure is present,

[11] As we shall discuss, it is not essential, although it is extremely helpful, for the class size treatment *per se* to have an impact on test scores on average in order for the identification strategy to work.

but this instrumental variables strategy of using the previous random assignment indicators as a forcing variable for the latent (or unobserved) lagged test scores is less susceptible to these specification problems than if the lagged test scores *were* observed, in which case more stringent identifying assumptions would have to be made.

The strategy then is to use the *contemporaneous* average of the peer group test scores $\bar{y}_{-i,j}$ as the peer group measure. The instrument for this measure, which tackles litany of endogeneity problems discussed above, is the fraction of the class net of student $i$ who were previously randomly assigned to a Small class:

$$z_{-i,j} \equiv \frac{1}{N-1} S_{-i,j} \tag{3}$$

with the analogous 'leave out $i$' quantity as:

$$S_{-i,j} \equiv \sum_{k \neq i}^{N} d_{kj} = S_j - d_{ij} \tag{4}$$

Note that this instrument handles the problem that the test scores for the New Entrants are not observed prior to their exposure to the treatment. In effect, we 'pick out' the component of the post-exposure test outcome that is due to having been exposed to the Small class treatment in the previous grade or not, and so use only that variation in the predicted peer group measure. The use of the lagged instrument also deals with the reflection problem, as only the component of the peer group measure that varies with the lagged treatment is used in the predicted peer group measure.

The presence of the current grade class type indicator $D_j$ in the regressor set, however, might render this nothing more than a conceptual discussion. In order for the instrument to have power, it must be that $z_{-i,j}$ be correlated with $\bar{y}_{-i,j}$ net of $D_j$. By the Frisch-Waugh Theorem, this means that $z_{-i,j}$, the fraction of student $i$'s classmates who were previously in Small classes, must have sufficient variation after its linear dependence on $D_j$ is factored out. This is clearly where the degree of New Entrants, and in particular, the extent to

which the New Entrants are spread across classes $j$ is key to give the instrument any chance of power in our data. As we show in Figures 1 and 2, fortunately for our purposes, the Fraction of New Entrants does indeed have significant variation across classes for all three grades. Figure 1 is a histogram of the fraction of each Small class who were previously randomly assigned to a Small class as well. Were there no new entrants, and no students switching class type, the histogram for each grade would be a single bar at 1. In fact, we can see while there is a pronounced tendency for that fraction to fall between 0.5 and 1, the histogram reveals substantial variability in this fraction across classes. Figure 2 does the same exercise for the Regular classes, where absent the new entrants and switchers, each histogram would be a single bar at 0. While the variation across classes here is less visually apparent, it is also clear we have some power. Finally, as we shall see below when we present the first-stage regression results, this net variation (net of the Small class indicator $D_j$) in the instrument also has decent explanatory power at the third grade level, and moderate at the second grade level, for the peer group outcomes.

The inclusion of the class type indicator $D_j$ also helps ease the exogeneity requirements for the group formation. For example, the presence of the class type indicator in the regression has the effect of sweeping out all observed and unobserved factors that vary purely at the class *type* level. So if we assume that the (possibly endogenous) sorting that takes place within class types of students and teachers into particular *classes* is the same for the Small and Regular classes, then the presence of the $D_j$ treatment indicator will 'balance the bias' (Heckman, 1997) and net it out of our estimated equation. The point is that randomization creates two groupings of students and teachers that are, in principle, identical on either side of the treatment and control line. While the sorting *within* the two clusters of students and teachers into classes may well be endogenous, as long as that process is the same for both groups, the presence of the treatment indicator will guarantee that it

will be differenced out. Of course, if students and teachers are assigned not just to class *types* on the basis of randomization, but also individual classes within class types, then this entire discussion is moot. But we have been unable to verify with certainty that all schools in the Project STAR experiment created classroom groupings via randomization, and so we proceed under these weaker assumptions. While the idea of identical endogenous processes leading to class formation (under the scenario where we dispense with the possibility that classes were formed via a randomization scheme), we should mention it is not difficult to construct behavioral models in which these processes would not be identical owing precisely to the differing class sizes on either side of the treatment and control lines.[12] That is a very nuanced version of the endogenous sorting story, and to speak more to it empirically would require far richer data than we have access to here.

Our instrumental variables strategy yields differences in the power to detect peer effects across grades. First, it should be obvious by the very nature of our identification strategy, in that it relies on the lagged Small class assignment variable, that peer effects will not even be estimable via this strategy for Kindergarten. Given that not all children attend Kindergarten in Tennessee, this is perhaps not a serious shortcoming of our strategy. By default, we assign all of the reduced form effect to the 'pure' class size effect in examining the Kindergarten class type estimate, although what we are really saying is that, given our identification strategy, we cannot *tell* if some portion of this effect is really being driven by peer group effects or some other source. Similarly, while we are not prohibited from empirically estimating a peer group effect for the First grade with our strategy, as we will see, we really have quite low empirical power. This brings us to the conceptual point we wish to make on this subject in this section. Because our identification strategy literally

[12]A point we owe to Andy Foster for pushing us to think beyond the purely statistical statement of this identifying assumption.

relies upon the *feedback* of the treatment assignment effect on students as the Project STAR cohort ages, we expect to see greater notional power for the later grades. We wish to stress that of course the failure to detect an effect does not imply there is *no* effect, and so in our context the failure to detect peer effects in the early grades may simply be symptomatic of the very design of our identification strategy.

To summarize this section, we rely most heavily on the aspect of Project STAR that it randomly assigns a Small class treatment to individuals and then clusters those children differently as the experiment progressed across grades. This is the key to our identification strategy in extracting measurement of the endogenous peer group effects from these type of data. We will discuss the specific econometric properties of our estimation scheme and how it fits in with a more general discussion of peer group effects in Section 5 below. We do not argue that the students in Project STAR are randomly placed into individual classes, but merely class types (Small or Regular) within each participating school. The technical literature on this aspect is unclear, and in any case, our strategy is operational if, as we assume, students and teachers are only guaranteed to be assigned randomly to class types and not purely classes. The bottom line is we are relying on the social multiplier effects of the class size reductions to identify the peer effects and not the random assignment of students to different peer groups. The extra assumption we must incur lacking the random assignment to individual classes is that the potential sorting that does occur along the lines of our instrument is the same process across the two randomly determined treatment groups. Finally, as we stated at the outset, we have for now adopted the canonical approach of the literature in other respects, such as adopting the regression model that is linear in the peer group mean outcome as well as the extremely critical assumption that the relevant peer group is the student's classmates as regards the test score outcomes.

# 4 The Evidence on the Social Multiplier Effects of the Small Class Size Treatment in Project STAR

In this section we use the Project STAR data together with our identification strategy just discussed in the previous section to estimate peer effects. Before we move to that estimation framework, we first replicate the earlier work done with Project STAR on the class size effects as in Krueger (1999), and then interpret these as reduced-form (or total) class size effects that we try to pull apart into their underlying components of the peer group effect and the residual which we call the 'pure' class size effect. We consider both the instrumental variables as well as the reduced form results, the latter of which combine the direct class size effects together with the social multiplier or feedback effects created by the experiment. The reduced form allows us to begin to perturb the canonical framework to alternative specifications. We also consider the robustness of our baseline instrumental variables results to alternative instrumentation strategies, as well as assess the sensitivity of our results to departures of the Project STAR data from the experimental protocol (such as class type switchers).

## 4.1 Estimates of the Peer Effects and the Pure Class Size Effects: Inside the Black Box of Project STAR

We begin our empirical analysis with first presenting the reduced-form class size effects using the Project STAR data. The results are broken out by the four grades for which the experiment ran, and as we discussed above, all regressions include school fixed-effects as the STAR data were not a random sample of schools. Owing to the randomization of students and teachers within schools to the three class types - Small, Regular, and Regular with a teacher's aide (we use Regular as our omitted base group) - a simple OLS

regression estimates the treatment effects of interest as the coefficients on the Small and Regular/aide dummies.[13] These results are presented in Table 3, and our results reproduce the analogous results presented by Krueger (1999) and Hanushek (1999) (without regard to their subsequent interpretation of these results). In short, the Regular/aide classes do marginally better, although the difference is not statistically distinguishable from the Regular class base group. The Small class estimates, however are all quite significant at conventional levels, and range from a low of 4.8 percentile points to a high of 7.3 percentile points relative to the Regular class students. It is not much violence to these results to summarize them as saying that being in a Small class appears to have roughly a 5 percentile point gain over students in Regular classes (of either type) at each of the four grade levels.

What we wish to do is essentially pry apart this 5 percentile effect into its constituent components of a pure class size effect and the peer group effect which is the focus of our work. An alternative statement of our goal is to split the class size effect into its direct and indirect effects, although this language is rather imprecise and leaves the implications for policy counterfactuals rather muddled. Whereas earlier authors, especially Krueger (1999), interpreted the roughly 5 percentile point gain implied by the coefficient on the Small Class indicator as pertaining to the causal effect of the Small Class *size* as compared to the omitted control group, Regular Classes, we wish to remain more agnostic at this stage.

We interpret this as the total effect of being assigned to the Small Class *type*, but we view this categorization as a bundle of components which comprise the 'black box' of the class type, and which may include peer effects and other elements of a general schooling production function. At the inception of the program (i.e. Kindergarten and possibly First

[13] In an experimental setting, the inclusion of covariates helps in countering small imperfections in the randomization along observable dimensions, but primarily serves to reduce the residual uncertainty and so reduce the sampling error of the effects of interest.

Grade) it seems plausible that the cohort design to the study would more precisely reflect a pure class size effect. But as the cohort ages, it becomes increasingly difficult to argue that the simple contrast between the Treatment and Control groups reflects a pure class size effect, without allowing for the possibility that the experimentally induced changes in the peer group compositions might also play a role. What the earlier literature as exemplified by Krueger (1999) and Hanushek (1999) focused on was the lack of *widening* of the gap between the students who remain in the Small classes as the experiment progressed, and why the 5 point gain appeared to be a once and for all gain, as opposed to an increase in the slope of the test score-grade relationship as well as in the intercept.

Table 4 presents the simplest possible departure from the Treatment and Control indicators used to measure the class size effects from Table 3. In Table 4 we include the additional characteristic of the classes given by the average (leave-out mean) test score of the class $\bar{y}_{-i,j}$ - a measure we intend to capture the 'peer group effects' as articulated in Section 3 above. We are not trying to ascribe any behavioral significance to these regressions, but we want to present a benchmark by which the IV estimates we present below might be compared. In particular, owing to the reflection problem which we discussed in Section 2, the individual outcome $y_{ij}$ and the peer group outcome $\bar{y}_{-i,j}$ are simultaneously determined and so the reverse causality would have to be considered formally to give this a behavioral interpretation.[14] The remarkable stability of the estimated coefficients across grades on the

[14]As we discussed in Sections 2 and 3, we do not have test score outcomes for the New Entrants prior to their entry to the Project STAR schools. Therefore, we cannot resort to *ad hoc* fixes to the reflection problem by utilizing a lagged version of the peer group measure (i.e. the student's current peers' test score in the *previous* grade). However, we did use, purely for comparison sake, the lagged mean peer group effect lagged one grade for those students who *were* in the Project STAR schools in the previous grade. This exercise has the effect of replacing the reflection problem which hinders the behavioral interpretation of the results in Table 4 with another problem, which is, what does the lagged peer group measure mean if it is only

peer effect measure certainly presage the analytical results we consider in the next section and in the Appendix that derive the sample properties of the type of peer group estimators considered in Table 4. Across the three columns, we see that the estimated coefficients on the peer group measures are virtually identical at 0.58 with standard errors of 0.04. The coefficients on the Small class indicators exhibit a little more heterogeneity across grades, and they have fallen to roughly half their original magnitudes from the total program effect estimates given in Table 3. The point estimates suggest a small decline in the Small class effects across the three grades (as in Table 3), although the decline is not statistically significant. All three estimates of the Small class effect, however, remain statistically distinct from zero even after including the contemporaneous peer effect measure as an additional regressor.[15]

At the bottom of Table 4 we present what we call the normalized peer effect which places the estimated coefficient on the peer group measure given in the first row of each

constructed over those students who were in the experiment last year? Interpretation problems aside, we find the biggest change occurs in the first grade, where the estimated coefficient on the peer effect drops from the estimated 0.58 in Table 4 to 0.05 with a standard error of 0.07. The second and third grade estimates on the peer group measure drop by about half to 0.21 for both grades. For the most part, the Small class dummy coefficients remain qualitatively the same, although the point estimates show a more pronounced monotonic decline across grades. But as both versions of Table 4 suffer from measurement or simultaneity problems, we only use them to serve as a benchmark to contrast our later results to.

[15]Here again we would be remiss if we did not point out the presence of the reflection problem and the problems with interpreting the results in Table 4 behaviorally. As regards the Small class effect, obviously one potential impact is that it enhances the performance of a student's peers. Therefore, including it as a covariate will obviously diminish the potential effect of the Class size mechanism, as it simply splits the total effect displayed in Table 3 into a direct and indirect effect, with the contemporaneous peer group measure being a potential outcome of the *contemporaneous* class type indicator. The IV estimators considered below do not have this mechanical problem of simply splitting the overall effect of purely the contemporaneous class size measure.

column on the same scale as the coefficient on the Small class indicator. Conceptually, it captures the discrete effect of moving from a Small to a Regular sized class on the average peer group measure. From a measurement perspective, we can view the sum of the effects on the Small class indicator and on this 'normalized' peer group effect as roughly splitting the overall (roughly 5 percentile point) reduced-form experimental effect into its constituent components of the direct class size effect and the feedback effect induced by the peer group effect. As we can see in the last row, the normalized peer effects reflect the homogeneity of the peer effect coefficients and they vary from roughly 4 to 3 points. If we sum the Small class effect in the second row of Table 4 with the normalized peer effect, we get the estimated *total* Small class effects of 6.66 for First grade, 5.26 for Second grade, and 4.95 for Third grade. If we compare these to the total experimental effects of the Small class type presented in Table 3, these were 7.31, 5.94, and 4.76. Thus, for the most part, the Small class direct effect and the normalized peer effect combined appear to account for the average total experimental effect of the Small class assignment.

We turn now to our instrumental variables strategy which avoids the reflection problem and also accounts for the aspect of the sampling design of the experiment in that we do not have test scores for the New Entrants prior to their joining the Project STAR schools. As we discussed in the previous section, we use as an instrument for the contemporaneous peer group measure $\bar{y}_{-i,j}$ the fraction of the current peer group students who were assigned to the Small class treatment in the previous grade $z_{-i,j} \equiv \frac{1}{N-1}\sum_{k\neq i}^{N} d_{kj}$. The instrument therefore treats students who are either New Entrants to the experiment or previously randomly assigned to one of the Regular class types as the same as far as explaining variation in the class to class variation in average test scores.[16]

[16]To the extent that the 'Regular' class size represents the average class size in the schools from which these students came, this may not be such a bad approximation. The random assignment of the New

As we noted in our conceptual discussion in the previous section, this strategy looks to have promise since the fraction of students who were previously randomly assigned to a Small class has good variation across classes for the Small class type group (owing to the significant quantity of the New Entrants). In Table 5 we present the first stage of the projection of $\bar{y}_{-i,j}$ on $z_{-i,j}$. We do this by grade, and as the grade increases, obviously the number of potential instruments grows, as students may have first been exposed to the Small class treatment in an ever-increasing number of prior grades. So, for example, by the third grade, there are three such possible instruments. By looking at the first three rows of Table 5, the reader can see that for the most part, the instruments are individually generally not statistically distinct from zero. The exceptions to this are the Kindergarten effect for the Second grade regression, which is marginally statistically significant, and the rather large point estimate for the Third grade, which is highly significant at conventional levels. The joint test on the combined significance of the instruments for each regression is given in the 4th row from the bottom of the table. There the reader can see we have quite low power for the First grade, weak to moderate power for the Second grade, and quite good power for the Third grade owing largely to the Kindergarten peer measure effect. This pattern of power for our instrumental variables framework we anticipated in our previous conceptual discussion of our strategy, as it relies directly on the feedback notion of what a peer group effect is, and so it only becomes detectable as the cohort ages and the feedback effects potentially surface from the environment.

Notice also that because we are instrumenting for a grouped version of the dependent

Entrants to the Small and Regular class types helps balance the differences between the New Entrants and the previously assigned students along unobserved dimensions once the contemporaneous class type indicator $D_j$ is conditioned on. As we noted in Section 2, however, there is plenty of evidence to suggest that *unconditionally* the New Entrants and those students previously randomly assigned to even just Regular classes *are* observationally distinct.

variable, the first-stage regression is also almost the reduced form for the two equation system at the individual level.[17] Therefore, we can also examine the effect of the class type indicators after holding constant the direct peer treatment effects of interest. This approach has the advantage of avoiding any sort of reflection type problems. However, as regards our principle identifying assumption, it may be subject to the endogenous sorting objection if the sorting is systematically different between the Small and Regular classes. But keeping with our assumption that this bias is balanced across the treatment arms of the experiment, then the coefficients on the Small class indicators tells us to what extent the Small class effect of Table 3 is only reflective of the spillover effects generated by the past impact of the experiment. Indeed, while the Small class effect for the First grade, 6.39 (and statistically distinct from zero), is still close to its Table 3 estimate, the point estimate for the Grade 2 effect is half its Table 3 value, and is statistically indistinguishable from zero. Finally, the Grade three point estimate is actually negative, but is again indistinguishable from zero. Thus, our conclusions which we shall discuss below regarding the insignificance of the Small class effects at Grade 2 and 3 of the Project STAR experiment are not subject to a criticism that we may have mishandled the treatment of the endogenous peer effects. Once measures capturing the prior exposure to the Small class treatment of an individual's peers are included, the current effects of having been assigned to a Small class are substantially

[17]The use of the term 'almost' here may be unclear. For the most part, the dependent variable in the first stage regression presented in Table 5, $\bar{y}_{-i,j}$ varies little across students within classes, but more so across classes. Below we shall consider reduced forms purely at the classroom level, as the treatments of interest vary only at the class level rather than the individual level, and so in this sense, the standard errors presented in Table 5 over-count the degrees of freedom for these treatments. The class level results are presented in Appendix Table 1, and show that our correction for the within-class correlation of the errors almost completely compensates for the possible overstatement of the degrees of freedom. Thus, inferences drawn from Table 5 are not deceptive owing to the 'over-counting' of the degrees of freedom.

attenuated.

The second stage instrumental variables results presented in Table 6 represent the core results of our paper. They show that once we account for the simultaneous determination of the individual $y_{ij}$ and contemporaneous peer group outcomes $\bar{y}_{-i,j}$ using the lagged fraction of the peer group exposed to the treatment as an instrument, the estimated peer effects swamp the direct Small class size effects in grades 2 and 3. The first grade point estimate of the peer effect is roughly one-third of the second and third grade estimates, and is quite imprecisely estimated. As such, it is indistinguishable from no effect, although as we indicated above, and we wish to stress again, this lack of finding an effect should certainly not be construed to imply that there is no effect, as the power of the empirical design is quite weak here. Indeed, the confidence interval on the first grade effect more than encompasses the Second and Third grade effects, and so could even be construed as consistent with those point estimates.

The normalized peer effects are presented in the last row of Table 6, and roughly speaking, the Second and Third grade effects have a point estimate of about 4.5. The Small class effects presented in the second row are now extremely small relative to the 5 percentile point estimates of the overall effect presented in Table 3, and quite indistinguishable from zero. Given the precision of the standard errors on these two point estimates, we can clearly reject their equality to the earlier reduced-form effects. This pattern is reversed, however, for the First grade estimates. There the Small class effect remains largely unchanged at 4.91, although the standard error on this estimate is extremely large, so it is also indistinguishable from zero. The estimated normalized peer effect is less than half the grade two and three effects, at roughly 2 percentile points. The associated $t$-statistic, however, is less than 0.30, reflecting the low power, and as we already noted, the peer effect for the First grade is indistinguishable from 0.

This very stark pattern of the apparent *complete* overtaking of the Small class size effect by the peer effect as of the second grade may strike the reader as unusual, and perhaps indicative of some spurious attribute of the setting driving these results. For that reason we next turn to examining the sensitivity of these basic results to alternative specifications and measurement schemes. However, it is also useful to pause for a moment and point out one exercise this paper will not be able to shed much light on. Namely, as a measurement device, we have posited that individual outcomes vary with the mean outcomes of the reference group. But we have not considered the behavioral model by which these individual outcomes, which are presumably the result of underlying choices and inputs, come to be influenced by the reference group. Manski (2000) among others has delineated three broad channels by which the peer group mechanism might propagate: 1. Preference interactions 2. Expectation interactions and 3. Constraint interactions. While we certainly agree that for the evidence in this paper to lead to precise policy prescriptions we would need to establish how these behavioral mechanisms lead to the peer group influences we observe, we emphasize that the Project STAR data do not sample characteristics that enable us to speak to these alternative explanations empirically. It is possible at this juncture to offer stories which might rationalize this pattern of results across grades that rely differently on say the preference versus the expectations rationales behind the peer influences, but we shall avoid this *ex post* theorizing in this paper, and leave this exploration until the relevant variables can be sampled.

## 4.2 Assessing the Robustness of the Peer Effect Results

Table 7 presents our first set of robustness checks of our basic specification presented in Table 6. Essentially this table is concerned with the fact that since each student in Project STAR can be represented as a given experimentally assigned 'type', then using one source

of variation is equivalent to using one minus another source of variation. For example, each student currently in a Small class was either previously randomly assigned to a Small class last year (PRASC), a New Entrant to the Project STAR schools (NE), or one of the rather small fraction of class type Switchers (S). If we let each of these variables denote their respective fractions, then we have for each Small class:

$$1 = PRASC + NE + S \tag{5}$$

So then it is identically true that for just the Small classes, using the fraction PRASC as an instrument, as we did in Table 6, is equivalent to using (1 - NE - S) as an instrument.

For the Regular type classes, a student who was PRASC who is now in a Regular class is clearly a Switcher, and so we will replace the designation of switcher to PRASC for the Regular classes, to keep the notation for a Switcher, S, as being *just* for those who switch from a Regular to a Small class. Introducing the notation of PRARC for those students who are in a Regular class now who were previously randomly assigned there, we have:

$$1 = PRARC + NE + PRASC \tag{6}$$

So now we have the identity that PRASC = 1 - NE - PRARC, and so using PRASC as an instrument for the Regular classes is identical to using 1 - NE - PRARC as an instrument. Notice we have purposefully not used notation to distinguish between New Entrants to Regular classes versus New Entrants to Small classes, as the randomization should equate those two groups. However, PRARC and PRASC are potentially distinct groups as they have been exposed to different treatments at an earlier point in the experiment.

The basic point of spelling out these identities is that using the variation explained by the proportions of students in the classes who were, for example, previously randomly assigned to a Small class is identically the same as using the 'mirror image' (and thus the same first stage projection and the same IV estimate) proportions of the other groups of

students across classes. This point is useful to keep in mind in interpreting Table 7. First, we can examine the possibility that the peer effect works differently for the Small and Regular class types. Therefore, the first row of Table 7 pools the class types as in Table 6 and uses PRASC as the instrument, thus replicating the first row of Table 6. The next two rows allow the peer effect to be potentially different across class types. For the Third grade, the estimated peer effect coefficients are roughly the same, and roughly average to the pooled Third grade effect presented in Table 6. For the Second grade, the Small class peer effect is roughly the same as the pooled Second grade effect from Table 6. However, when looking just within the Regular classes, the estimated peer effect is highly imprecise and the point estimate is actually negative. Now here is where the identities just presented become useful. As we noted above, for the Regular classes, the number of students PRASC is equal to the number of Switchers (into Regular class types). Therefore, a regression which uses only the fraction of class type Switchers will produce an identical point estimate, and by looking at the last row of Table 7, the reader can see that the -0.56 point estimate from the third row is identical to the -0.56 estimate for the Second grade in the last row.

Thus, when we allow the peer effect coefficient to differ by class type, we can see that in the case of the Second grade, the point estimate is quite different for the Regular classes than for the pooled (across class types) estimate given in Table 6. Likewise for the peer group effect for the Regular classes for the First grade as is shown in the first column of the third row of Table 7. In contrast to the Table 6 pooled estimate, the point estimate here is roughly the same magnitude (and statistical significance) of the Second and Third grade estimates from Table 6. And of course here again, the estimate is identical to the First grade estimate for the Regular classes in the last row of Table 7 which uses the fraction of class type Switchers as the excluded instrument. Our point in displaying this numerical equality of the estimated effects, as well as the brief conceptual discussion we just provided

on the 'reverse image' form of identification is precisely to highlight to a skeptical reader that our identification strategy uses different groups to identify effects when we pool across class types. The reader, for possibly good reasons, may be worried about relying *entirely* on class type switchers to identify a peer group effect, as students who opt to switch class types (in this case the somewhat more unusual choice of switching from a Small to a Regular sized class) is endogenously determined with respect to the outcome. Such readers may therefore wish to discard those aspects of our analysis that include these Switchers as a source of identifying information. For this reason, they may wish to instead focus on the Small class estimates given in the second row of Table 7, as opposed to the pooled class type estimates given in Table 6.

The Small class peer effect estimates from the second row of Table 7 are qualitatively the same as the pooled peer effect estimates from Table 6 for the Second and Third grades. The First grade peer effect estimate, while still quite imprecisely estimated, is now quite large at 1.72 and is statistically distinct from zero at conventional levels. This discrepancy with our Table 6 results is in some sense reflective of the low power properties of our identification design with regards to the First grade setting that we have discussed previously. Across the multiplicity of specifications we have presented both in the paper, as well as those not presented, we tend to find much more systematic and uniform peer effect estimates for the Second and Third grade, whereas the results for the First grade are much more mixed and far more specification dependent.

This pattern is also seen in the specifications we present in the middle rows of Table 7 where we now use the percent of New Entrants in the class as the instrument for the peer group measure. The idea here is to use the variation in peer group 'quality' induced by those students who were *not* exposed to either the treatment or control groups of Project STAR. As we noted in discussing our primary identification strategy underlying Table 6,

we might expect that the New Entrants are comparable to the students already assigned to the control classes in the Project STAR groups, but if there is some type of spillover, or simply that the New Entrants represent a distinct group apart from the pre-existing Project STAR students, then this strategy might be appropriate. Our primary intent, however, is not to offer a strong behavioral justification for this instrumental variables strategy, but simply an alternative measurement strategy of the peer group coefficient. For the most part, our conclusions from the other parts of Tables 6 and 7 stand. The Second and Third grade results tend to be statistically distinct from 0, although the estimated effects are diminished in comparison to Table 6. This is especially true when we break the estimated effects out by class type and we look at the effects for just the Regular classes - these effects are roughly half of their Table 6 counterparts. Part of this attenuation might arise from the mixing of the students previously assigned to *either* Small or Regular classes under this identification strategy. For the First grade, we do find a statistically and economically significant estimated effect for the pooled class type specification, but the effect estimated for just the Small class types is highly imprecise, and for the Regular class type it is just below conventional levels of statistical significance. Overall, this alternative measurement strategy does not alter our primary conclusions from Table 6, and this is especially so as we think more carefully as to *what* source of variation this alternative strategy picks out of the variation in peer quality across classes.

Finally, we present what might be thought of as the 'perverse' source of variation in peer quality across classes, and that is using the fraction of students in each class who opt to switch away from their initially randomly assigned class type. As we discussed in Section two, and was also discussed in Krueger (1999), this may not be such a contaminated source of variation as the reader might think at first blush, as many of the students who switch class types are documented to do so for disciplinary reasons and the like. Thus, it is not obvious

that a student who switches from a Regular to a Small class does so because she is more academically motivated. For that matter, we should mention that while the numbers are only about one-third as large, we do see some degree of switchers in the opposite direction, from Small to Regular classes. For the average student, it may be plausible to think that these different groups of switchers, from Regular to Small and from Small to Regular, impart different biases on the estimated effects, which is why we present them broken out by class type (and not pooled) in the last two rows of Table 7. For the reasons just discussed, therefore, it is perhaps not too surprising that the point estimates of the peer effects based on those who switch into the Small classes from the Regular classes (presented in the next to last row) are slightly larger than the estimates based on the switchers from the Small to Regular classes (presented in the last row). However, the differences are extremely slight (a maximum of 0.07) and are not statistically significant across grades. Thus, if the reader does posit that the switchers endogenously select into class sizes based on preferences for academic 'quality' we find no statistical evidence of a systematic bias in one direction based on these estimates. For that reason, we have not excluded the switchers from our overall analysis as these estimates and further specification checks indicate that they do not exert a systematic influence on our estimated effects. We interpret this as confirming the Project STAR informal survey-based evidence and Krueger's (1999) evidence that the treatment of switchers in alternative specifications is essentially inconsequential in the impact on the final results.

### 4.3 What Do the Peer Effects Mean?

Our intent so far has been to take the canonical approach of estimating an equation such as:

$$y_{ij} = \beta \bar{y}_{-i,j} + x'_{ij}\gamma + \epsilon_{ij} \tag{7}$$

and to construct useful estimates of the peer effect manifested in $\beta$ by utilizing the random assignment features available in the Project STAR data. In particular, our identifying strategy relied on the notion that subjecting a student to the Small class type treatment induced not just potentially a boost in that child's test score outcome, but an indirect or spillover effect on the child's classmates through the peer group effect. This is what we mean by the feedback or social multiplier effect of the Small class type treatment. However, as has been noted frequently in the literature, the linear-in-the-peer-group-mean specification just presented implies that *given* a population of students of a particular quality, a reallocation of those students into alternative groupings would lead to the same aggregate outcome if this specification accords to the underlying mechanism generating the data.[18] We now turn to the question of whether, *given* the Project STAR treatment, non-linearities in the peer group effect exist so that reallocations or alternative groupings of students exposed to the treatment affect aggregate output. In terms of policy questions, this would speak to the pure efficiency implications of 'ability tracking' in which classes are formed to homogenize along the basis of initial test score outcomes.

To examine this, we turn directly to the class-level reduced forms (as instrumenting non-linear versions of $\bar{y}_{-i,j}$ obscures the basic point) where we allow the instrument of the percentage of students previously randomly assigned to a Small class to enter in a rather arbitrarily non-linear way by breaking the percentage into five dummies as it varies from 0 to 100 percent. We have included the other covariates in these specifications by grade (including, of course, class type) but have suppressed reporting those coefficients for

[18]But just to re-emphasize, it is *not* true that this implies that the class size treatments applied to a population of $NJ$ students individually produces the same aggregate output compared to the design of it being applied to $J$ groups of $N$ students each. The latter design contains the feedback or social multiplier effect of the Small class type assignment we are attempting to measure. If this is not clear at present, we hope that it will be clear to the reader by the end of the next section.

brevity. Unfortunately, as the relevant variation here occurs at the class level and we have only about 330 classes in the data, we have little power to detect these non-linearities. This is compounded by the fact that for each grade, only a little more than 100 of the classes (of either the Small or Regular type) contain more than 20 percent of children who were previously randomly assigned to a Small class. The average cell size outside of this base group, therefore, is only about 30 classes. For the most part, the linear-in-the-group-mean model appears to be consistent with the data. There is extremely slight evidence of a larger point effect once the fraction of students who were previously in a Small class passes a 40 percent threshold for all three grades. And there is also slight evidence in the Third grade of a larger benefit as this threshold is moved to 60 percent. Assuming a particular parametric form of the non-linearity would lend greater power to this exercise, but we were unable to quantify a convincing non-linear pattern that we felt appropriately summarized this reduced form. Such non-linearities may exist, but it will likely take a sample much larger than the Project STAR design in the number of classes dimension to measure them with accuracy.

In Table 9 we take on the idea that it may not be the 'quality' of a student's peers that matters for individual outcomes, but more of the 'sameness'. That is, imagine a school in which an entire first grade class is promoted intact to the second grade, so that the student's classmates remain exactly the same. In Table 8 a class in a cell like '80 to 100 percent of classmates were previously randomly assigned to a Small class' might have simply been a Small class that was moved virtually intact across grades. Looking at Table 8 we cannot tell if the estimate was created by the 'sameness' of the class, or because the class was exposed to the Project STAR Small class treatment. In Table 9 we include an additional set of dummy controls, analogous to those used in Table 8, to control for an arbitrary non-linear profile of class 'sameness' - i.e. the fraction of the class that was previously in the same class

together. Interestingly, even with this additional set of controls for class 'sameness', the conclusions of Table 8, with only a slight non-linearity appearing in the Third grade at the 60 percent threshold, appear to hold up quite well. One feature that might be interesting for future work on this topic is that the class 'sameness' estimates tend to be larger than the 'quality' estimates for the Second grade estimates. However, the opposite is true for the Third grade estimates where the 'quality' or Small class treatment exposure measures tend to have estimated coefficients which are larger than the 'sameness' coefficients.

# 5 Sample Properties of the Peer Group Effects and Alternative Estimation Schemes

Until now, we have asked the reader to bear with the canonical regression-based estimation framework of extracting peer group effect estimates from a sample of data. We have argued that the Project STAR data provide a superior means of estimating such effects because it uses randomization to allocate individuals to treatment and control groups, and these individuals are sampled over time so that the resulting feedback, or social multiplier, effects of the social program can be extracted from the data. That framework consists of the (appropriately instrumented) '$y$ on $\bar{y}$' regression familiar from studies in the literature that try to get at quantifying endogenous peer group effects. We turn now to 'unwrapping' this '$y$ on $\bar{y}$' regression by working out its properties *in the sample*. Much work has been done on the conceptual and population aspects of the peer effects model, but very little has been done on spelling out exactly what sample information is being used to produce an estimated effect. We show that our instrumented peer effects model employed in the previous section in fact captures the very essence of an endogenous peer effect, that being the social multiplier or feedback effect, of the social program used to create the instrument. We then relate our

approach to other innovations in the empirical study of externalities, as well as recall related discussions from the early union wage effect literature on the differing effects estimated by individual-level and industry-level data which pertain to spillover effects.

The '$y$ on $\bar{y}$' approach makes sense from the usual perspective of trying to quantify a relationship where an outcome of interest is regressed on an input or regressor of interest (generally net of other covariates, but this is unimportant to the ideas considered here.) However appealing though that might be, this regression also comes very close to running a regression of $y$ on itself - the $y$'s being for other individuals in the sample being the only aspect saving this from being purely tautological. Least squares estimators have the property of placing the fitted regression line through the point of means of the dependent and independent variables of the regression. Therefore, even when the regression is not literally a regression of $y$ on the $y$ for the same individual, a coefficient of 1 may still be produced purely because least squares is the estimating procedure - it tells us nothing about the underlying true parameter values generating the data.

In fact, we show in the Appendix the relevant algebra that establishes the sample properties of several estimation schemes in which the estimator equals 1 without considering any underlying data generating process. The first of these is the OLS case when the group mean *inclusive* of individual $i$ is used as the regressor, for example because the data sample only a fraction of the hypothesized peer group (such as the entire school in the High School and Beyond or the National Education Longitudinal Study).[19] However, of more relevance to our work is the Instrumental Variables estimator where the instrument is the full group mean (again, *inclusive* of individual $i$), but the peer group measure is the 'leave out mean' as we are able to use with the Project STAR data. This estimator also provides a sample

[19]Altonji (1988) considers alternative estimation schemes for group characteristics when the sample contains only a small fraction of the relevant group members.

estimate of 1 regardless of the underlying data generating process.

The empirical literature on peer effects has been especially pre-occupied with tackling the endogenous peer group affiliation problem. For that reason, the recent papers by Zimmerman (1999) and Sacerdote (2001) which use the random assignment conventions of a few colleges in designating freshmen roommates have drawn some appeal. As we discuss below, however, relying purely on random group assignment to study peer effects leaves the researcher an estimator that is still rather 'fragile' in its properties. The point of this paper, however, is that access to a randomized social experiment, whereby a treatment alters the outcomes of some of the individuals and the peer group formation is the same process across the experimental groups, allows for estimators which are not as fragile in extracting meaningful peer group estimates from the data. To put this more succinctly, the presence of a randomized social experiment of varying intensities across groups allows the researcher to *directly* investigate the presence of spillover effects. We present the relevant derivations behind this argument now, and then see how they tie-in to the instrumented peer group regression methods we utilized in the previous section. We then conclude with a general discussion of the estimation of spillover or externality effects from other literatures.

We begin with a stripped-down version of our estimating equation (leaving out covariates for the moment, dropping considerations of timing of the outcome and peer group measure, and assuming the group sizes are of homogeneous size N):

$$y_{ij} = \pi \bar{y}_{-i,j} + v_{ij} \tag{8}$$

In general, even in the absence of covariates, this regression will not produce a coefficient of 1, unlike the 'full mean' specification discussed in the Appendix when no covariates were included. Re-writing the 'leave out mean' in terms of the full group mean and the individual

outcome, we have:

$$\bar{y}_{-i,j} = \frac{1}{N-1}(N\bar{y}_j - y_{ij}) \tag{9}$$

Therefore, the OLS estimator for the regression just given is:

$$\hat{\pi} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{N}[\frac{1}{N-1}(N\bar{y}_j - y_{ij})y_{ij}]}{\sum_{j=1}^{J}\sum_{i=1}^{N}[\frac{1}{N-1}(N\bar{y}_j - y_{ij})]^2} \tag{10}$$

Simplifying this, we have:

$$\hat{\pi} = \frac{(N-1)\sum_{j=1}^{J}[N^2(\bar{y}_j)^2 - \sum_{i=1}^{N}(y_{ij})^2]}{\sum_{j=1}^{J}[(N^3(\bar{y}_j)^2 - 2N^2(\bar{y}_j)^2 + \sum_{i=1}^{N}(y_{ij})^2]} \tag{11}$$

Now, since $\sum_{j=1}^{J} N(\bar{y}_j)^2$ is simply the Between Sum of Squares (BSS) in the outcome variable and $\sum_{j=1}^{J}\sum_{i=1}^{N}(y_{ij})^2$ is the Total Sum of Squares (TSS), we may write this expression in the more interpretative form using this notation:

$$\hat{\pi} = \frac{(N-1)[N \cdot BSS - TSS]}{N(N-1)BSS - (N \cdot BSS - TSS)} \tag{12}$$

Finally, using the notation WSS for the Within Sum of Squares, and making use of the equation $TSS = BSS + WSS$, we can rewrite this as:

$$\hat{\pi} = \frac{BSS - \frac{WSS}{N-1}}{BSS + \frac{WSS}{(N-1)^2}} \tag{13}$$

We can use this last expression to begin to develop some intuition for the least squares '$y$ on $\bar{y}$' regression by unwrapping how it utilizes variation in the outcome measure within and between groups. First, notice that this OLS estimator of the peer group effect goes to 1 'mechanically' (i.e. regardless of the underlying true value of the peer effect) as one of two things happen: (i) The reference group size $N$ goes to infinity and (ii) the Within Sum of Squares (WSS) in the outcome measure goes to 0. This tells us immediately that our sample will have no power to detect (true) peer effects if there is no variation in the outcome measure within groups but only across groups. This would occur, for example, if groups were constructed by ability grouping used in schools where variation in student ability

occurs mostly across classes rather than within classes. Failure to account for institutions and behavioral mechanisms that lead to the formation of homogeneous groupings along reference group lines can easily lead the researcher to spuriously conclude peer effects are present. Similarly, the ideal data contain a large number of reference groups so that the reference group size is not too large relative to the overall sample size, and $N$ does not grow at too fast a rate as the overall sample size increases.[20]

By ignoring the term in the denominator that is down-weighted by order $N^2$, we can derive a more intuitive expression that approximates equation (13):

$$\hat{\pi} \approx 1 - \frac{WSS}{(N-1)BSS} \tag{14}$$

This expression is key to our 'unwrapping' of the '$y_{ij}$ on $\bar{y}_{-i,j}$' regression. Simply put, if reference groups are literally the sum of their parts then there are no spillover or peer group effects. Consider altering individual $i$'s outcome in a peer group of size $N-1$ (i.e. net of individual $i$ herself). If the resultant increase in the $WSS$ is exactly $(N-1)BSS$, i.e. the blip in the within-group variation *only* shows up in the between-group variation appropriately 'inflated' by the net group size $N-1$, then the estimated peer effect will be zero. If, however, the between group variation increases by *more* than the $N-1$ contribution from individual $i$'s impact on the within-group variation, then the estimated peer effect will be greater than zero. The upper bound on the coefficient estimated via OLS is 1, which occurs when the variation in individual outcomes is purely across groups rather than within groups.

Equation (14) is the key to our following analysis. It illustrates the basic intuition that the between group variation in outcomes contains the spillover (or peer) effects, whereas the

[20]Power considerations, which we do not examine here, would place a brake on driving the optimal reference group size too close to zero, as does the tradeoff in reducing the Within Sum of Squares as the group size diminishes.

within group variation gives a 'clean shot' of the individual variation purged of the group-level peer effect. The same principle that group level versus individual level data on the same variable contain different spillover or sorting effects is also the basic principle underlying the identification strategies in Boozer (2001) and Senesky (2000), both of whom use contrasts within and between groups to purge or extract effects which manifest themselves purely at the group level. Of course, the idea is not new, as the work of Lewis (1963, 1987) on union wage effects articulated this point carefully. In Lewis's case, the early industry level data on unionization percentages and average wages of workers contained not only the direct impact of (individual) union status on wages, but also the potential 'union threat' mechanism whereby higher unionization percentages in an industry meant the union could extract greater demands in the form of wages. Thus, Lewis viewed the 'union threat' effect as a nuisance and a possible reason why the early estimates based on aggregate data might overstate the individual union wage effect based on micro data.

The Lewis 'threat effect' corresponds to our peer group effect. In our setting it is actually the object of interest as opposed to a bias that needs to somehow be eliminated. The analytics given above lay out how the two forms of estimating 'the' union wage effect - via aggregate-level or individual-level (micro) data - combine to estimate the full set of parameters. As we just discussed, were we interested solely in the *direct* effect, we could utilize the purely within-group individual-level variation to estimate an effect purged of the spillover or peer group effect. Of course, to make the analogy to Lewis more exact, we need to introduce the analogous variable to his unionization status which in our case would be class size. Before coming to the specific treatment of dealing with class size, let us start by adding covariates to the simplified regression given in equation (8).

In this case, we amend equation (8) as:

$$y_{ij} = \beta \bar{y}_{-i,j} + x'_{ij}\gamma + e_{ij} \tag{15}$$

In this case, a simple application of the Frisch-Waugh Theorem allows us to apply the intuitive approximation we derived in equation (14) to the variation in the outcome net of its linear dependence on the covariates $x'_{ij}$ denoted as:

$$\hat{\beta} \approx 1 - \frac{(WSS|x'_{ij} - \bar{x}'_j)}{(N-1)(BSS|\bar{x}'_j)} \tag{16}$$

where the overbars denote the sample means of the respective variables. This expression highlights the sensitivity of the estimated peer effects to the *type* of covariates included in the regression. For example, a covariate that varies solely at the group or classroom level, such as teacher characteristics or the current class type, affects only the between variation in BSS. It has no effect on the conditional WSS as it is orthogonal (by construction) to the WSS. Therefore, adding a covariate that varies solely at the classroom level *unambiguously* drives down the estimated peer effect, the more so as the covariate is related to the cross group variation in outcomes. This is an alternative statement of the 'reflection problem' in that all characteristics of the common environment shared by individual $i$ and her peers must be controlled for, or the estimated peer effect will be overstated.

Adding covariates that vary both within and between groups or classes, such as student race or gender, have an ambiguous effect on the estimated peer effect. Their effect depends on whether they explain relatively more of the within or the between class variation in test scores. To the extent that they largely soak up the within class variation, but less of the between class variation, this will lead to a larger estimated peer effect that approaches 1. A covariate that affects the within and between variation 'proportionately' (i.e. a 1 unit change in a covariate for the within variation equates to a $\frac{1}{N-1}$ unit change in the between variation for a given individual) will contribute zero to the estimated peer effect, as no spillover is present.

Studies which rely purely on exogenous (or randomly formed) group assignment mecha-

nisms, such as Zimmerman (2000) or Sacerdote (2000), essentially follow the approach just described. They include in the covariates a number of factors which describe the individual heterogeneity, and run a regression of the individual outcome on a lagged version of the outcome of their randomly assigned college roommate. The discussion we just presented shows that their estimated effect relies entirely on how the covariates affect the variation in outcomes within and between roommate pairs. If the covariates do little to control for the possibly heterogeneous environments shared by roommate pairs in the between pair variation, but they parse out individual variation quite well, then such studies may be estimating spuriously large peer group effects. As we show in the Appendix, the lagging of the outcome variable (to overcome the simultaneity problem) used as the key right-hand side regressor simply modifies the expression given in equation (16) by multiplying it by the autocorrelation coefficient in the current and lagged outcomes being used in the regression. If the randomization of the roommates is done correctly, and the appropriate covariates are controlled for, then our observations here do not indicate a specific problem with such studies. However, we do wish to point out the 'fragile' nature of the identification achieved by relying solely on exogenous group formation, and the sensitivity of such estimates to the inclusion and exclusion of potential covariates. In addition, as we discuss in the next subsection, the use of random assignment for group formation has the problem that for large enough group sizes $N$, the variability in peer composition *across* groups goes to zero as $N$ increases. Thus while randomization helps ensure group formation is exogenous, it runs the risk in large group settings that the peer effect will not even be identified. In small groups, the variability across groups will arise due to the finite-$N$ sampling error.

We turn next, therefore, to the empirical strategy we have used in this paper. This does not rely on randomized group assignment as in Zimmerman (1999) and Sacerdote (2001), but instead on the hypothesis that *conditional on the current class type assignment* $D_j$, the

treatment status in the earlier grade, $d_{ij}$, of a student's peers is exogenous. The inclusion of the current class type dummy $D_j$ in the list of covariates allows that if there *is* endogenous selection into individual classes based on the $d_{ij}$'s of the class, it must be the same process for both the Small and Regular classes, so that the bias is thus differenced out across the treatment and control lines by the presence of $D_j$. The necessary exclusion restriction is that another student's (call them $k$) prior treatment status $d_{kj}$ has no impact on student $i$'s outcome *except* via the endogenous peer effect mechanism. Thus, we take the instrument for the endogenous $\bar{y}_{-i,j}$ to be:

$$z_{-i,j} \equiv \frac{1}{N-1} \sum_{k \neq i}^{N} d_{kj} \tag{17}$$

And as above, since the reference group size $N$ is taken as constant across groups, define the part of the instrument $z_{-i,j}$ that varies by $j$ as:

$$S_{-i,j} \equiv \sum_{k \neq i}^{N} d_{kj} \equiv S_j - d_{ij} \tag{18}$$

with $S_j$ being simply the total number of students previously assigned to the Small class treatment in the current class $j$. Finally, in order for the instrument to have power *conditional on the covariates* (most importantly, conditional on the class type indicator $D_j$) we need to assume the assignment status to the Small class previously has an effect above and beyond the current class type status. Simply put, this means we need the Small class assignment to have not purely just a once and for all effect, but also an effect on the *slope* of the test score profile across grades. In fact, empirically we come dangerously close to not having any power, as Krueger (1999) reports that much of the Project STAR effects are of the once-and-for-all variety. However, he also presents point estimates that show a slope effect that is about one-fifth the size of the 5 percentile point 'intercept' effect. Thus, while the power is reduced it is still present, and it is worth noting, the power will also tend to be greater the *earlier* in the experiment the student was assigned to the Small class treatment,

for this reason.

With this instrument in hand, we now consider the sample properties of the Instrumental Variables estimator of equation (15), where the peer group measure $\bar{y}_{-i,j}$ is taken to be endogenous and instrumented with $\bar{z}_{-i,j}$. Taking again the simplification that the group size $N$ is the same across groups, the IV estimator is:

$$\hat{\beta} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{N}\frac{1}{N-1}S_{-i,j}y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{N}\frac{1}{N-1}S_{-i,j}\bar{y}_{-i,j}} \tag{19}$$

Again, the $N-1$ factor divides out of the numerator and denominator and this simplifies to:

$$\hat{\beta} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{N}(S_j - d_{ij})y_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{N}(S_j - d_{ij})[\frac{1}{N-1}(N\bar{y}_j - y_{ij}]} \tag{20}$$

and multiplying out and passing the sum over individuals through the numerator and denominator yields:

$$\hat{\beta} = \frac{(N-1)\sum_{j=1}^{J}[NS_j\bar{y}_j - \sum_{i=1}^{N}d_{ij}y_{ij}]}{\sum_{j=1}^{J}[N^2S_j\bar{y}_j - 2NS_j\bar{y}_j + \sum_{i=1}^{N}y_{ij}d_{ij}]} \tag{21}$$

Now make use of the same notation, BSS, WSS, and TSS (to refer to the Between, Within, and Total Sum of Squares respectively) as above, but here applied to *covariances* between the outcome and treatment indicators, rather than pure variances in the outcome variable within and between groups (just for economy of notation in this step). Recalling our notation that $S_j = Nz_j$, we again have:

$$\hat{\beta} = \frac{(N-1)[N \cdot BSS - TSS]}{N(N-1)BSS - (N \cdot BSS - TSS)} \tag{22}$$

which is the same expression, in terms of sums of squares, that we had above in equation (12). Using the operator $Cov$ to refer to the sample covariance, it again simplifies down to be approximately:

$$\hat{\beta} \approx 1 - \frac{Cov[(y_{ij} - \bar{y}_j), (d_{ij} - \bar{d}_j)|x'_{ij} - \bar{x}'_j]}{(N-1)(Cov(\bar{y}_j, \bar{d}_j|\bar{x}'_j)} \tag{23}$$

What is somewhat more comforting about this expression than the analogous expression given in equation (16) for the randomized-groups peer effects estimator, is that it relies not just on the univariate variation in the outcome (net of the covariates) within and between groups, but instead now relies on the co-variation in the outcome with the previous treatment assignment dummy $d_{ij}$ within and between groups. Then, to the extent that the covariation is larger Between classes than Within classes, the second term will be driven to a quantity less than 1, and a positive estimate of a peer effect will result. Whereas the pure random-assignment OLS estimator in (16) relies crucially on both the randomization being done properly as well as (more importantly) the type and quantity of the covariates which are included, the IV estimator properties just spelled out in equation (23) indicate that the IV estimator is less fragile to the specification and takes advantage of a randomly allocated program at the individual level.

However, the spurious detection of peer group effects may still arise in the IV case. If, contrary to our assumptions, prior treatment assignment $d_{ij}$ is used as a factor in assigning students to classes, and in particular such that there is no within-class variation in $d_{ij}$, then in general we will estimate a spurious peer effect of 1. What we require, therefore, is that students are assigned to individual classes, *conditional* on their current class type $D_j$, such that $\bar{d}_{-i,j}$ is an exogenous variable.[21] In the context of Project STAR, this requires that to the extent the New Entrants are placed into individual classes in a way that is related to their outcome variable differently than those students who were in Project STAR from

[21]Clearly, unconditional on $D_j$ this is certainly not the case. Owing to the experimental design, students who remained in the Project STAR schools from grade to grade remained in the same class type, apart from the small number of switchers. Therefore, overall, students who were in a Small Project STAR class last year are *much* more likely to be found in a Small Project STAR class this year. The question of exogeneity, therefore, is if students are clustered into individual classes *within class type* in a manner systematically related to $\bar{d}_{-i,j}$.

the previous grade, then this differential assignment mechanism must be the same for the Small and Regular classes. So either (i) there is no endogenous sorting on the basis of the treatment assignment $d_{ij}$ into classes, or (ii) to the extent there is endogenous sorting, the 'bias is balanced' across the Treatment and Control groups.

By inspection, equation (23) hints that the instrumented '$y$ on $\bar{y}$' regression coefficient may be estimated *without* placing the outcomes of one's peers as a regressor on the right-hand side. Instead, when a social program is available, then an appropriate comparison of the ratio of the effects of that social program within and between classes can provide evidence of endogenous social effects *without* resorting to the rather uncomfortable '$y$ on $\bar{y}$' device. We take up this analysis in the next subsection. The discussion also shows the ties of the endogenous peer effects literature to other similar estimators of spillover or externality effects, the linkages to which have not been entirely clear in the existing literature.

## 5.1 Alternative Estimation Schemes to the Canonical Approach Based on Within and Between Group Contrasts

We begin this subsection by comparing the tradeoffs between the random peer assignment strategies utilized by Zimmerman (1999) and Sacerdote (2001) to the 'social program' strategy used in this paper of identifying peer effects. The first thing to note in the random assignment case is that the estimate of the peer group effect is generally heavily over-identified. The reason is that to the extent that individual outcomes are influenced by observable characteristics such as gender, race, family background, etc. and the group compositions vary along these observable lines, then the peer effect can be estimated off these varying group compositions. For each observed characteristic of the individual a separate peer effect can be estimated, provided the variation in the individual characteristic across groups is sufficient. If the researcher maintains the hypothesis that the peer influences work

through the outcomes (i.e. the endogenous effects model of Manski) then the empirical model will be heavily overidentified. Of course, one quirk of relying on the random group assignment hypothesis is that as the group size $N$ tends towards infinity, the variation in group characteristics will tend to zero if indeed groups are formed via a randomization scheme. In finite group sizes, there will tend to be variation in characteristics across groups due to sampling error. For this reason, the college roommate context considered by Zimmerman (1999) and Sacerdote (2001) where $N$ is quite small (generally 2 or 3) is ideal. But one should be careful in considering asymptotic properties of estimators under the random group formation hypothesis, in that only the number of groups be allowed to approach infinity and not the group size. In the latter case, the model would be asymptotically unidentified.

We also consider in this subsection a weaker identifying assumption that pertains to our Project STAR data. In that case, the classes themselves are not necessarily randomly formed, but only the class types. However, we argue in this paper that classes are exogenously formed along the lines of the fraction of child Previously Randomly Assigned to a Small Class *conditional* on the class type indicator (as well as the other covariates). In that case, we can no longer rely on the demographic or individual characteristics to provide a source of necessarily exogenous variation in peer qualities, but only have the experimentally induced variation arising from having been previously exposed to the Small class treatment in the Project STAR schools. Thus we lose the overidentified nature of the empirical model with the gain of allowing for weaker identifying assumptions.

We are going to use equations 14 and 16 as the intuitive basis that a moments estimator constructed from the Within and Between class estimators of the Previously Randomly Assigned to a Small class indicator (PRASC, denoted above as $d_{ij}$) will replicate the instrumental variables estimator of the '$y$ on $\bar{y}$' peer effects regression. This is the same idea

pursued in Boozer (2000) whereby IV estimators based on group-level characteristics can be seen as contrast or moment estimators based on how the stochastic processes vary within versus between groups. In the present context, this has a direct analogy to the early work of Lewis (1963, 1987) regarding what aggregate or industry level data versus individual level data on unionization identifies. This also has the effect of linking our analysis to concepts relating to the peer effect, such as Philipson's (2000) 'external treatment effect' which measures the spillover which may arise in medical vaccination trials, whereby greater density of vaccination may have larger aggregate benefits, even holding constant the total number of vaccinations administered. Finally, in the case where the analyst, like Lewis in dealing with the 'union threat effect', finds the spillover or externality effect a nuisance parameter, the estimation scheme discussed below allows for a pure estimate of the *direct* Small class size effect, net of the peer effect feedback.

Rather than do the tedious algebra to show the estimator we propose is numerically identical in the sample, we choose the simpler task of showing that they have the same limiting value as the sample size grows due to the number of groups growing large, holding class sizes fixed. We first pose the endogenous peer effects data generating process (*dgp*) as:

$$y_{ij} = \delta d_{ij} + \beta \bar{y}_{-i,j} + \theta D_j + x'_{ij}\rho + u_{ij} \tag{24}$$

Notationally, $d_{ij}$ indicates if the child was previously randomly assigned to a small class, and $D_j$ indicates if the current class is Small or not, and so it has no within-class variation for a given class indexed by $j$. Since the sample average of $\bar{y}_{-i,j}$ to the class level is simply $\bar{y}_j$, the Within class estimator derived from applying OLS to the following regression (with the $f_j$ denoting the class specific fixed effects):

$$y_{ij} = \alpha d_{ij} + \lambda \bar{y}_{-i,j} + x'_{ij}\kappa + f_j + e_{ij} \tag{25}$$

can be written in terms of the *dgp* (dropping the error terms for ease of exposition) as:

$$y_{ij} - \bar{y}_j = \delta(d_{ij} - \bar{d}_j) + \beta(\bar{y}_{-i,j} - \bar{y}_j) + (x'_{ij} - \bar{x}'_j)\rho \tag{26}$$

Then, using equation 9, the term involving the peer effect can be simplified to:

$$y_{ij} - \bar{y}_j = \delta(d_{ij} - \bar{d}_j) - \frac{\beta}{N-1}(y_{ij} - \bar{y}_j) + (x'_{ij} - \bar{x}'_j)\rho \tag{27}$$

And so the Within class regression of individual test scores on the previously randomly assigned to a small class last year dummy as well as the individual-level covariates will estimate, in terms of the *dgp*:

$$y_{ij} - \bar{y}_j = \frac{\delta}{1 + \frac{\beta}{N-1}}(d_{ij} - \bar{d}_j) + (x'_{ij} - \bar{x}'_j)\frac{\rho}{1 + \frac{\beta}{N-1}} \tag{28}$$

As the group size $N$ is large, then the within-class estimates will come very close to delivering a clean shot of the *direct* effect of having previously been randomly assigned to a small class. As the magnitude of the peer effect in our case is less than 1, but the group size is roughly 20, we can almost safely ignore this 'correction' to the within estimates of delivering a clean shot of the direct effect of the prior experimental status purged of the feedback spillover effects. However, when the group size is roughly 2 or 3, as in the case of Zimmerman (1999) or Sacerdote (2001) who study college roommates, this correction is less likely to be negligible. The correction arises because, when the peer groups are small, each individual's contribution to the peer effect is non-negligible. In that case, the within-class regression will tend to *understate* the direct effect because the within regression subtracts out part of the direct effect by netting out the group mean in $\bar{y}_j$.

Similarly, we can examine the limiting properties of the Between class regression, where OLS is applied to the class averaged data:

$$\bar{y}_j = \psi\bar{d}_j + \bar{x}'_j\tau + \omega D_j + \nu_j \tag{29}$$

Again, ignoring the true error term, we can re-write this in terms of the parameters of the *dgp* as:

$$\bar{y}_j = \frac{\delta}{1-\beta}\bar{d}_j + \bar{x}'_j\frac{\rho}{1-\beta} + \theta(1-\beta)D_j \tag{30}$$

Therefore, if we focus on the Within and Between class estimators of the coefficients on the $d_{ij}$ PRASC indicator, we have that the Within estimator has the limit (limits being taken as $J$, the number of groups, tends to infinity):

$$plim\,\hat{\alpha} = \frac{\delta}{1+\frac{\beta}{N-1}} \tag{31}$$

and similarly, the effect on $\bar{d}_j$ in the Between estimator has the limit:

$$plim\,\hat{\psi} = \frac{\delta}{1-\beta} \tag{32}$$

Thus, to a first approximation, for the class size $N$ large, we can form an estimator for the peer effect $\beta$ as:

$$plim\,(1-\frac{\hat{\alpha}}{\hat{\psi}}) \approx \beta \tag{33}$$

The intuition is that the Within estimator $\hat{\alpha}$ estimates the direct effect of PRASC purged of the class-level peer effect due to the inclusion of the $J$ class dummies. On the other hand, the Between estimator $\hat{\psi}$ will estimate an 'inflated' version of the direct effect, which is inflated the more that the peer effect $\beta$ tends towards 1. In the case where the Within and Between estimates of the PRASC effect are the same, the implied peer effect is therefore zero. But in large samples, the Between class estimate of PRASC will tend to be larger than the Within class estimate. In this setup, however, nothing about the construction of the estimator implies the estimated peer effect from a finite sample will be bounded on the interval from 0 to 1.

Of course, since the group size $N$ is known (and in the analytics here, assumed to be constant across groups, unlike in the Project STAR data where it varies slightly, thus

introducing another form of approximation) we can provide the exact minimum distance estimator based on the Within and Between estimators as:

$$plim\left[(1-\frac{\hat{\alpha}}{\hat{\psi}})(\frac{N-1}{N-1+\frac{\hat{\alpha}}{\psi}})\right]=\beta \tag{34}$$

which is slightly attenuated for large $N$ from the approximate form we gave above. Also notice that as the fraction $\frac{\hat{\alpha}}{\hat{\psi}}$ goes to 0 (i.e. the implied peer effect goes to 1) the approximation also becomes exact. Roughly speaking, if we take the ratio of the Within to the Between estimates to be 0.5, and $N=21$, this correction shows up only in the second decimal place, and is thus well within the sampling error of our estimates of the peer effects in the previous section. Similarly, the exact estimator for the direct effect of $d_{ij}$ is not simply the Within estimator $\hat{\alpha}$, but instead a slightly larger version:

$$plim\left[\hat{\alpha}\left(\frac{N}{N+(\frac{\hat{\alpha}}{\psi}-1)}\right)\right]=\delta \tag{35}$$

Here again, in the case where there is no spillover effect manifested in the estimates, the Within and Between estimates will be the same, and so indeed the Within group estimate will be an estimate of the direct effect of the Small class size effect purged of the group level feedback effect. And even in the presence of a feedback effect, for large enough group sizes $N$, the Within group estimator of the treatment effect provides a clean estimate of the direct effect of the program, net of the social multiplier effects. Of course, identification requires that the fraction of those treated vary within groups (and groups are not segregated by treatment status, as would be the Project STAR data were their no New Entrants and perfect adherence to the experimental design protocol) as well as that fraction must vary *across* groups so there is variation in the $x$ variable of interest.

Next we turn to the important observation that in studies where a randomization device is used to assign peer groups, the implied peer group effect will generally be overidentified. The reason is that often the researcher has available other characteristics of the individuals,

captured in the regressors $x'_{ij}$, that are associated with differences in student performance. As such, even though there is not a social program *altering* individual performance as is the case with the PRASC indicator $d_{ij}$, the differing compositions of peer groups as reflected by $\bar{x}'_j$ allow for identification of the peer effect coefficient $\beta$ by contrasting the Within and Between coefficients on the $x$'s in equations 21 and 23 in the manner just discussed above for the regressor $d_{ij}$. Take for example the $k$th element of the coefficient vectors on the $x$s from the Within regression in equation 18 and the Between regression in equation 22, then we should have for each element in the regressor set that:

$$plim\left[(1-\frac{\hat{\kappa}_k}{\hat{\tau}_k})\left(\frac{N-1}{N-1+\frac{\hat{\kappa}_k}{\hat{\tau}_k}}\right)\right] = \beta = plim\left[(1-\frac{\hat{\kappa}'_k}{\hat{\tau}'_k})\left(\frac{N-1}{N-1+\frac{\hat{\kappa}'_k}{\hat{\tau}'_k}}\right)\right] \qquad (36)$$

thus showing the overidentified nature of the random group formation case when the analyst has information on more than one individual characteristic that varies in intensity across groups. The caveat here is that as $N$ gets large, then if groups are truly formed randomly, the variance in the cross-group variation in average group characteristics will shrink to zero. For finite $N$, there generally will be variation in the averages that arises due to sampling error. Thus, ideally the analyst will have access to data in which the average group size in the randomly formed groups case is small, as otherwise the ability to detect peer effects will be minimized. In this respect, the college roommate setting of Zimmerman (2000) and Sacerdote (2000) is ideal, as $N$ is quite small. In Project STAR this would be more of a problem were classroom assignments, rather than class type assignments, randomly determined as this would undermine the identification of peer group effects.

The points that we wish to emphasize from the discussion in this section are: (i) The linear peer group model that is typically used in the literature when groups are randomly formed is generally overidentified, as long as there remains sufficient variation in the exogenous characteristics across groups. This will tend to occur when the group size $N$ is small,

and the variation in group characteristics thus arises by sampling error - clearly, these characteristics must vary sufficiently across groups, and must be correlated with individual performance to be of value. The overidentification arises from the number of restrictions the randomization of group formation implies. (ii) Even in the absence of randomly formed groups, an exogenously assigned social program operating at the individual level will allow for identification of endogenous peer effects as long as the intensity of the program varies within and between groups. If the program varied only between groups, but groups were stratified by program status, then we could not separately identify the individual effect from the spillover effect created by the endogenous peer effects. Similarly, as the within group variation essentially only identifies the *direct* effect of the program, lack of variation in the fraction of participants in the social program across groups would eliminate the very source of variation that is crucial in identifying the feedback effects. This would arise in our context if students were placed in classes (and not just class *types*) randomly and class sizes were sufficiently large so as to eliminate variability in class characteristics which are needed to create differential exposures to the peer 'qualities'.

(iii) The fact that our Instrumental Variables estimator of the peer group effect can be derived as approximately 1 minus the ratio of the Within class estimator of the Previously Randomly Assigned to a Small class indicator (PRASC, or $d_{ij}$) to the Between class estimator, shows the tight relationship of the canonical peer group estimation scheme and other problems in applied work. Lewis (1963, 1987) noted in his work the tendency of the Between industry union wage effects to be larger than the micro data union wage effects (Within industry or not), and he carefully considered the possibility of a 'union threat' effect which is analogous to our peer effect spillover which was responsible for the wedge between these two sets of estimates. More recently, Philipson (2000) has proposed a framework to consider the extrapolation of individually based clinical trials for medical treatments, which

have varying levels of intensity in the treatment populations across sites. He points out that in the case of vaccinations, say, a spillover or externality arises when larger fractions of children are vaccinated for an unvaccinated child. He proposes random assignment of treatment status intensities not only within sites, as is classically done in clinical trials, but *between* sites so as to allow for assessment of what he calls the 'external effects'. Such a two-stage randomization design, he argues, allows for extrapolation of the micro level clinical trials to a macro level setting by handling explicitly the 'implementation bias' that arises because of the external effects. In fact, by comparing his proposed estimators with the analytics we just presented - in particular, the equivalence of the IV-endogenous peer group effect approach with the 'contrast' estimator based on the ratio of the Within and Between estimators - the reader can see that, conceptually at least, his proposed estimation scheme is our 'unwrapped' endogenous peer effect estimator using the exogenously assigned social program $d_{ij}$ as the driving force behind the peer group 'quality'.

## 5.2 Empirical Results Based on the Within and Between Class Comparison of Prior Treatment Status Effects

In this subsection we make use of the within and between class relations between the prior Small class treatment assignment variable $d_{ij}$ and individual test scores $y_{ij}$. We focus our empirical work here on illustrating equations (24) to (33) in the previous section using the Project STAR data. In Table 10 we present in the upper panel the between class estimates of the current class type ($D_j$) effect, as well as the fraction of the class previously randomly assigned to a Small class $\bar{d}_j$. As the number of classes is roughly five percent of the total individual-level sample, the standard errors are quite large. The grade one Small class effect is now slightly larger than in Table 6, for example, at 6.48, and it statistically significant with a wide confidence interval. The grade two effect is indistinguishable from zero, and

the point estimate is roughly half the reduced-form 5 percentile point grade two effect. The point estimate for the grade three effect is actually negative, although is statistically indistinguishable from zero.

Now as equation (32) shows, the estimates of the coefficient on $\bar{d}_j$ across classes will be an 'inflated' version of the direct effect of $d_{ij}$ on student performance as long as the peer effect $\beta$ is greater than zero. For the first grade, the between class estimate of the effect of $\bar{d}_j$ is 1.53, and is indistinguishable from zero. For grade two, the effect is somewhat larger at 4.26, but is again well within sampling error of zero. For grade three, however, we see a quite large estimated effect of 13.77 with an associated $t$-statistic of over 3.

The within class estimates in the bottom panel are more precisely estimated owing to the larger degrees of freedom. As we showed in equation (31), for a group size of roughly $N = 20$, the within class estimates of the coefficient on $d_{ij}$ is essentially the direct effect of this variable on student performance purged of the feedback or peer effects. The deviating from class means of the covariates also eliminates the current class type $D_j$ as a regressor as it varies only across classes. In contrast to the role played by $\bar{d}_j$ in explaining the cross-class variation in the top panel, in the bottom panel, the largest estimated effect of $d_{ij}$ occurs for the first grade. The estimate there is 3.64, which is statistically distinct from zero, but statistically indistinguishable from the reduced form Small class effect in Table 3. The second grade estimate of the direct effect of $d_{ij}$ is 1.53 and it is well within sampling error of zero. While this within class estimate is not statistically distinct from the corresponding between-class estimate of 4.26 in the top panel, it is roughly one-third the size of the between class effect, suggesting a role for a spillover (or peer) effect at the class level. Finally, the grade three direct effect estimate is 2.33 and is statistically distinct from zero at conventional levels. However, as the cross-group effect in the upper panel is so large at 13.77, then this is rather strong evidence of a spillover/feedback effect at the group level.

In the last row of Table 10 we have computed the implied point estimate of the peer effect $\beta$ (in equation (24)) using equation (33). While we have not yet computed the delta-method standard errors, it should be clear to the reader the peer effects estimated this way will have a *much* wider confidence interval than the corresponding peer effect estimates computed via IV in Table 6. The implied grade one effect is actually negative, although it is clearly quite imprecise and so well within sampling error of a zero effect, consistent with the 0.3 (and statistically insignificant) estimate in the first row of Table 6. The grade two estimate of 0.64 is well within sampling error of the 0.86 peer effect estimated via IV in Table 6. We should note, however, that as the between class estimate of the grade two effect of 4.26 is statistically non-distinct from zero, the implied peer effect computed via equation (33) is likely not distinct from zero either, as it is the between estimate that contains the information on the spillover effect. Finally, we see roughly the same result for the implied grade three effect of 0.83, which is quite similar to the corresponding effect from Table 6 of 0.92.

In Figure 3, we have plotted the third grade within and between class relations between $y_{ij}$ and $d_{ij}$ net of the other covariates (notably the current class type $D_j$) via the Frisch-Waugh Theorem. We have super-imposed the relations on top of the between class Frisch-Waugh residuals for the 322 classes (the within class data points being far too numerous to display meaningfully). This plot shows that there is not just a cluster of classes or individuals driving these estimated relationships, but the effect is spread throughout all 322 classes. The fitted regression lines show the larger gradient for the between class relationship as compared to the within class relationship, thus yielding visually apparent evidence of a spillover effect via equation (33). The two lines cross at the point where $\bar{d}_j$ and $\bar{y}_j$ net of the covariates is 0. As these are fitted (Frisch-Waugh) residuals, this is the overall sample mean of both $d_{ij}$ and $\bar{d}_j$ by the construction of the residuals.

# 6 Conclusions

There has been a recent spate of exciting new empirical work documenting the existence and magnitude of peer effects in educational and social settings generally. Some of this work has made innovative use of institutional rules which pair college freshmen in a randomized fashion with roommates, as in Zimmerman (1999) and Sacerdote (2001), thereby hurdling one large obstacle in this literature, that being the endogenous sorting of individuals into their peer groups. Of course, the random assignment itself solves only one of the many problems, well delineated by Manski (1993), that have plagued the advancement of this literature. Peer affiliation, issues of model specification such as timing and measurement issues generally, must still be pushed to the back burner even with such data. Furthermore, as we document in this paper, and Sacerdote (2001) notes in his work, random assignment alone does not allow for distinguishing what may be 'endogenous' peer effects - whereby an individual is directly affected by the *outcomes* of her peers, leading to a social multiplier or feedback effect - from 'exogenous' effects, whereby the individual is affected not by outcomes of her peers *per se*, but the characteristics of her peers.[22]

In this paper, we take this literature to the next step by making use of data with a social program administered in a randomized fashion at the individual level. The Project STAR data on the effects of class size reductions for early-elementary school students from Tennessee in the early 1980's is a very natural dataset to use for such a purpose. Owing to the cohort design of the experiment, as the cohort progressed from Kindergarten to the final

[22]In a recent paper, Moffitt (2001) corroborates this argument that merely doing random assignment of group memberships does not guarantee identification of the structural peer effects from the estimated reduced form effects if exogenous effects are allowed for in the *dgp*. Furthermore, he verifies our argument that a randomly allocated social program identifies endogenous spillover effects via a classical simultaneous equations framework for the $N = 2$ case. See the discussion surrounding his equation (10).

grade of the project. Third Grade, the exit and replacement of students out of and into the Project STAR schools provides a sample of classrooms with differing past exposure to the Small class treatment. If a social multiplier, or endogenous peer effect, is indeed present, then classes with higher intensities of students exposed to the Small class treatment in the past, should have a classroom-level effect that *exceeds* the individual-level effect by a margin greater than the share of students treated. In this way, data which contain a randomly allocated social program can measure a spillover effect of a social program *directly*, thereby assuring a finding of an endogenous peer effect. Data which consist of purely random pairings of students, with no social program present, must rely on more stringent identifying assumptions to make such a claim. Furthermore, we also show that experimental designs such as that proposed by Philipson (2000) to study the spillover or 'external' effects of medicinal trials are in fact the same notion as an endogenous peer effect, as his conceptual idea focuses on measuring the feedback effect of a clinical trial. In addition, he proposes a two-stage randomization scheme, whereby intensities of a clinical trial are randomly assigned not just to individuals within a site, but also what fraction of each site is eligible to receive the treatment. Such a design would be a welcome addition to social experiments more generally.

The question of endogenous peer effects or exogenous peer effects is highly important. Even apart from considerations on the cost side of a social program - especially factors such as fixed setup costs per locale, for example - the presence of endogenous peer effects on the benefits side implies an economy of scale. In that case, social programs which are clustered in nature will have greater benefits than those programs which are sprinkled across the landscape. In the context of education, this literature fits well with the research on the pure resource effects as it speaks to the efficient allocation of such resources within and between schools.

In this paper, our 'introduction' of the presence of the peer effects lurking in the reduced form Small class size effects of Project STAR turns out to have rendered the class size resource effects *per se* to a much smaller magnitude by grades 2 and 3. Our evidence implies that especially by grade 3, the 5 percentile point impact of the Small class treatment is almost entirely due to the feedback effect of the enhanced peer qualities due to the treatments in the earlier grades. The evidence on the grade 2 effect is less sharp, although it does appear that across various specifications, about half of the 5 percentile reduced form effect is attributable to the peer feedback effect. For Grade 1, we should re-emphasize the nature of our identification strategy implies we have much less power to detect a feedback effect at the early stages of the experiment. With that in mind, we find no evidence of an appreciable feedback effect for the first grade, and so attribute all of the 7 percentile reduced form effect to the Small class reduction *per se*. We do the same for Kindergarten, although that derives purely from the design of our identification strategy. In summary, our results imply that alternative policy structures, such as the tracking of children following the grade 2 and 3 patterns of Project STAR *without* the Small class reduction, would be expected to produce a similar set of outcomes from those derived by Project STAR. The peer effects themselves appear to have 'overtaken' the pure resource effects in the later grades of the experiment. That said, it is important to stress that we rely on the experimental assignment to the Small class to produce a 'boost' in achievement in order for our peer effect identification strategy. We do not read our results as implying class size reductions have no effect, but we offer a more in depth investigation of the mechanisms by which such resource alterations do have effects than have been offered by prior examinations of the Project STAR data. Such a re-interpretation is a by-product of our interest in utilizing the experimental STAR data to identify endogenous peer effects.

While the Project STAR data do offer some important advancements for the empirical

study of peer effects, it is important to note several of the pitfalls cited by Manski (1993) have been held outside the scope of this paper. Foremost is our assumption that the relevant peer group is the Project STAR classroom in the current year. The problem is that lacking such a strong assumption imposed on the empirical work, making headway with these data is virtually impossible. In the context of Project STAR, however, anecdotal and introspective evidence suggests that early elementary classrooms do exert a powerful influence, more powerful than any other readily identified peer group delineation observable with our data. In that sense we are as comfortable as we can be about this assumption with these data, and are rather fortuitous in having the elementary school setting as the context for our data. As Manski discusses, absent such an assumption of this type, the identification problem for the peer model is essentially insurmountable. The second hurdle we have avoided altogether in this paper is attempting to categorize the peer effects we do find into *how* they manifest themselves. Manski (1993) offers three such categorizations: (i) preference interactions (ii) constraint interactions and (iii) expectations interactions. The theoretical work by Lazear (2000), for example, is related to the constraint interaction category. The model proposed by Akerlof (1997) might be thought of as a mix of both preference and expectation interactions. Distinguishing between such models does appear to matter greatly for the structure of policies designed to capture the peer effect spillovers. However, the Project STAR data, while quite good at allowing the measurement of the spillover effects, samples little that would help us empirically distinguish between these alternative models of *how* the peer effects manifest themselves. We hope the results of this paper push researchers to turn their attention to empirically distinguishing between these alternative models of the underlying mechanisms.

# 7 Appendix: The Algebra of Instrumental Variables Estimation of the Endogenous Peer Effects Model

In this Appendix, we derive the properties of the instrumental variables estimator for the empirical endogenous peer effects regression where the researcher uses characteristics of the *full* group in the sample as either an instrument or regressor (i.e. both the IV and OLS cases). For simplicity of exposition, we ignore the presence of other covariates. Conditioning everything on a set of exogenous covariates $x'_{ij}$ does not change anything conceptually, although including them may in fact mask some of the 'mechanical' problems with either IV or OLS that we address here.

To start, consider the empirical specification for the endogenous effects model as:

$$y_{ij} = \bar{y}_{-i,j}\beta + \epsilon_{ij} \tag{37}$$

where the notation $\bar{y}_{-i,j}$ is the 'leave-out mean' of the test scores for classroom $j$. It is related to the usual sample mean of the of the class test scores (denoted as $\bar{y}_j$) by:

$$\bar{y}_{-i,j} \equiv \frac{1}{N_j - 1} \sum_{k \neq i}^{N_j - 1} y_{kj} \tag{38}$$

so that:

$$\bar{y}_{-i,j} = \frac{1}{N_j - 1}(N_j \bar{y}_j - y_{ij}) \tag{39}$$

In the case where the sample of the peer group includes the *entire* peer group (which, purely by assumption, we assume to be the student's immediate classmates), then it makes sense to relate student $i$'s outcome to the outcomes of the students in the class *other than* student $i$, hence the use of the 'leave-out mean' as the relevant peer group measure.[23] The

[23] In contrast, when the data contain only a *sample* of the peer group members, then use of the ordinary sample mean $\bar{y}_j$ is sensible. This is because individual $i$ is representative of other members of the class who may not have been included in the sample.

instrument that we propose in this paper to extract the exogenous variation in the peer group measures $\bar{y}_{-i,j}$ is the fraction of the class previously randomly assigned to a Small class. We do not include an additional subscript for the timing of the variables simply because that is irrelevant to this discussion. Let $d_{ij}$ be a dummy variable indicator for whether the student was previously a member of a Small class. Then our instrumental variable is given by:

$$z_j \equiv \frac{1}{n_j} \sum_{i=1}^{N_j} d_{ij} \tag{40}$$

A rewrite of the expression for the instrumental variable $z_j$, the usefulness of which will be apparent below, is:

$$z_j = \frac{1}{N_j} S_j \tag{41}$$

where $S_j \equiv \sum_{i=1}^{N_j} d_{ij}$ is simply notation for the number of students in each class previously randomly assigned to a Small class. This rewrite is useful because since this is controlled experimental data, the *number* of students in each class, $N_j$, essentially does not vary across classes, and so the $j$ subscript is superfluous. The variation in the instrument $z_j$ therefore comes entirely from variation in $S_j$ across classes - i.e. $z_j = \frac{1}{N} S_j$.

The instrumental variables estimator for $\beta$ in the empirical model given above is, for a sample of $NJ$ students in $J$ classes is just:

$$\hat{\beta} = \frac{\sum_{j=1}^{J} \sum_{i=1}^{N} S_j y_{ij}}{\sum_{j=1}^{J} \sum_{i=1}^{N} S_j \bar{y}_{-i,j}} \tag{42}$$

(The number of students in each class, $N$, simply divides out of both the numerator and the denominator.) Now we can make use of our relation of the leave-out mean to the usual sample mean to re-write this as:

$$\hat{\beta} = \frac{\sum_{j=1}^{J} \sum_{i=1}^{N} S_j y_{ij}}{\sum_{j=1}^{J} \sum_{i=1}^{N} S_j [\frac{1}{N-1}(N\bar{y}_j - y_{ij})]} \tag{43}$$

And now note that the only quantities left which are affected by the sum over the $i$ subscripts are only the $y_{ij}$ terms in the numerator and denominator, and so carrying those sums through, this simplifies to:

$$\hat{\beta} = \frac{\sum_{j=1}^{J} S_j \bar{y}_j}{\sum_{j=1}^{J} S_j [\frac{1}{N-1}(N\bar{y}_j - \bar{y}_j)]} \tag{44}$$

This expression is easily seen to equal 1 in the absence of other covariates. Notice this is not an asymptotic expression, but holds *in the sample.*

Furthermore, this algebra for the IV case shows that a coefficient of 1 will also appear in the OLS case where the *full* group mean, $\bar{y}_j$ is used as the peer group measure, a coefficient of 1. That this is true can readily be seen by inspection of equation (42), replacing $S_j$ with $\bar{y}_j$. The fact that both variables vary only at the group level $j$ implies the same algebraic simplifications will hold, and the equation analogous to (44) will again be 1.

Of course, since the setup just discussed delivers a coefficient of exactly 1, it is improbable a researcher would not realize his error, and opt for a different estimation strategy. In this sense, the addition of covariates $x'_{ij}$ may mask this issue to the researcher, as now the coefficient will no longer be exactly 1 in the general case. Assuming that at least some elements of the vector vary at the individual (as well as the peer group) level, the OLS estimator is now (using matrix forms, and $M_x$ being the idempotent projector into the subspace orthogonal to the space spanned by the columns of the $X$ matrix, $P_B$ being the idempotent matrix which averages to the group ($j$) level):

$$\hat{\beta} = [y' P_B M_x P_B y]^{-1} y' P_B M_x y \tag{45}$$

If we assume, for exposition only, that the regressor vector consists of only a signle non-constant regressor $x_{ij}$, then some straightforward but tedious manipulation allows us to

write this as:

$$\hat{\beta} = 1 - [(y'P_B y) \cdot (x'x)^{-1} - (y'P_B x)^2]^{-1}(y'P_B x)(y'Qx) \tag{46}$$

where the idempotent matrix $Q$ is the within (class) operator. Thus, in order for the numerator of the second term to be non-zero, the regressor $x$ must vary within and between class, as well as being correlated with the outcome $y$ in both dimensions in the sample. If the regressor varies only at the group level (in our context, this could be a teacher characteristic, for example) then again, the sample estimate of the peer effect will be purely 1.

Note however, that now the reasons for why the coefficient deviates from 1 are not entirely meaningless. Intuitively, the more the regressor $x$ explains the within group variation in the outcome as compared to the between group variation, the coefficient will be driven towards zero. In fact, substantially more simplification on the expression above tells us the estimated peer effect will attain zero when the following expression holds:

$$\hat{\beta}_{yx}(R^2_{yP_B x}) = \hat{\beta}^B_{yx} \tag{47}$$

In other words, when the OLS coefficient from a regression of $y$ on $x$ within and between groups down-weighted by the R-squared from a regression of $y$ on $x$ between groups (i.e. the squared sample correlation between the Between group variation in $y$ and $x$) equals the OLS coefficient obtained from the between group regression of $y$ on $x$. As long as the Between regression coefficient $\hat{\beta}^B_{yx}$ lies above this, however, the estimated peer effect will be non-zero. As we discussed in the context of the 'leave out mean' estimators used in this paper, intuitively this is because the covariate $x$ influences the cross-group variation in the outcome $y$ than would be expected than if there were no 'feedback' effect of the covariate $x$ creating a spillover at the group (class) level as compared to its effect at the individual-level, appropriately down-weighted.

## 7.1 The IV Estimator When the Peer Measure is Lagged

Since $y$ and $\bar{y}$ are determined simultaneously, some researchers (e.g. Zimmerman (1999) and Sacerdote (2001) among others) have posited instead that the influence of one's peers depends on their outcomes from some earlier period, and thus estimate a modified regression of the one given above as:

$$y_{ij,t} = \beta^L \bar{y}_{ij,t-1} + x'_{ij}\gamma + \epsilon_{ij} \tag{48}$$

where the subscripts $t$ and $t-1$ denote the period for the individual outcome and the peer effect respectively (the dating of the other variables is not essential to this discussion and so omitted for simplicity). To cut down on the clutter of notation, assume that the sample correlation between $y_{ij,t}$ and $y_{ij,t-1}$ net of the regressor is the same at the individual and the group level and represented by $\rho$. If we let the estimator for the lagged peer effect be denoted as $\hat{\beta}^L$, then comparing this estimator to the one based on the contemporaneous peer measure (i.e. $\hat{\beta}$) we have that:

$$\hat{\beta}^L = \rho\hat{\beta} \tag{49}$$

In other words, the peer estimator which is derived from a regression equation using a lagged peer measure uses the same information as the one derived from an equation using the contemporaneous measure, except it is 'corrected' by the autocovariance properties in test scores. But this estimator is just as inherently fragile as the one based the contemporaneous peer measure, but will mask the tendency to estimate a coefficient near 1, due to the down-weighting by the first-order autocorrelation coefficient estimate of test scores.

# References


[1] Akerlof, George A., (1997), 'Social Distance and Social Decisions,' *Econometrica*, 65 (5), pp. 1005-1027.

[2] Altonji, Joseph G., (1988), 'The Effects of Family Background and School Characteristics on Education and Labor Market Outcomes,' Working Paper, Center for Urban Affairs, Northwestern University.

[3] Boozer, Michael A., (2000), 'Identification of Structural Parameters in Data With a Group Structure: Using Alternative Comparisons and Understanding Their Coherence,' mimeo, Yale University.

[4] Conley, Timothy and Christopher Udry, (2000), 'Learning About a New Technology: Pineapple in Ghana,' Economic Growth Center Discussion Paper 817, Yale University.

[5] Finn, Jeremy D., and Charles M. Achilles, (1990), 'Answers and Questions About Class Size: A Statewide Experiment,' *American Educational Research Journal*, 27 (3), pp. 557-577.

[6] Folger, John, (1989), 'Editor's Introduction: Project STAR and Class Size Policy,' *Peabody Journal of Education*, 67 (1), pp. 1-16.

[7] Hanushek, Eric, (1998), 'The Evidence on Class Size,' Occasional Paper No. 98-1, W. Allen Wallis Institute of Political Economy, University of Rochester.

[8] Hanushek, Eric, (1999), 'Some Findings From an Independent Investigation of the Tennessee STAR Experiment and From Other Investigations of Class Size Effects,' *Educational Evaluation and Policy Analysis*, 21 (2), pp. 143-164.

[9] Heckman, James J., (1992), 'Randomization and Social Policy Evaluation,' in *Evaluating Welfare and Training Programs*, Charles F. Manski and Irwin Garfinkel, eds., Harvard University Press.

[10] Heckman, James J., Jeffrey Smith, and Nancy Clements, (1997), 'Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts,' *Review of Economic Studies*, 64 (4), pp. 487-535.

[11] Krueger, Alan B., (1999), 'Experimental Estimates of Education Production Functions,' *Quarterly Journal of Economics*, 114 (2), pp. 497-532.

[12] Krueger, Alan B., and Diane M. Whitmore, (2001), 'The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR,' *The Economic Journal*, 111 (1), pp. 1-28.

[13] Lazear, Edward P., (1999), 'Educational Production,' *NBER Working Paper 7349*.

[14] Manski, Charles F., (1993), 'Identification of Endogenous Social Effects: The Reflection Problem,' *The Review of Economic Studies*, 60, pp. 531-542.

[15] Manski, Charles F., (1995), *Identification Problems in the Social Sciences*, Harvard University Press.

[16] Manski, Charles F., (2000), 'Economic Analysis of Social Interactions,' *The Journal of Economic Perspectives*, 14 (3), pp. 115-136.

[17] Moffitt, Robert A., (2001), 'Policy Interventions, Low-Level Equilibria and Social Interactions,' in *Social Dynamics*, Steven Durlauf and Peyton Young, editors, MIT Press.

[18] Philipson, Tomas J., (2000), 'External Treatment Effects and Program Implementation Bias,' *NBER Technical Working Paper 250*.

[19] Sacerdote, Bruce, (2001), 'Peer Effects with Random Assignment: Results for Dartmouth Roommates,' *Quarterly Journal of Economics*, 116, pp. 681-704.

[20] Senesky, Sarah E., (2000), 'Commuting Time as a Measure of Employment Costs,' mimeo, University of California, Irvine.

[21] Word, Elizabeth, and John Johnston, et al. (1990), *The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Final Summary Report 1985-1990*,' Tennessee State Department of Education.

[22] Zimmerman, David J., (1999), 'The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR,' Working Paper, Williams College.

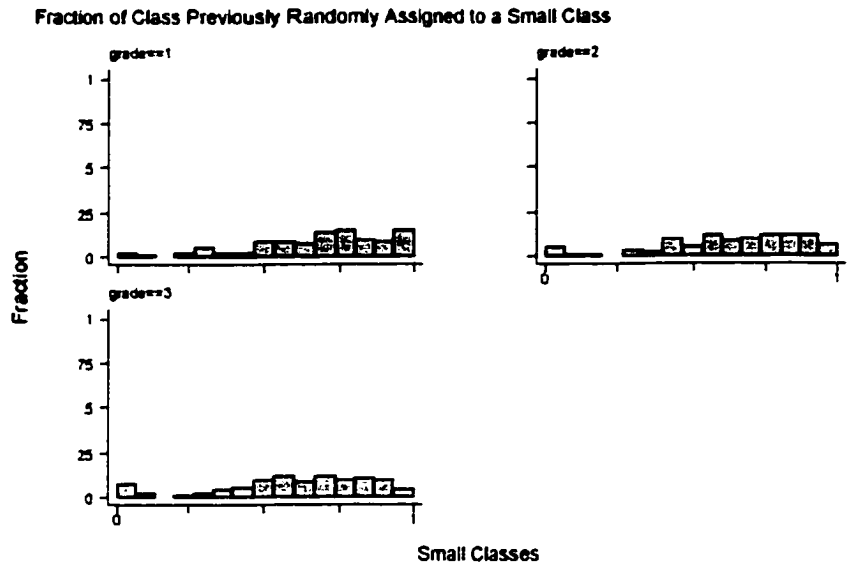


Figure 1: Class Level Histograms

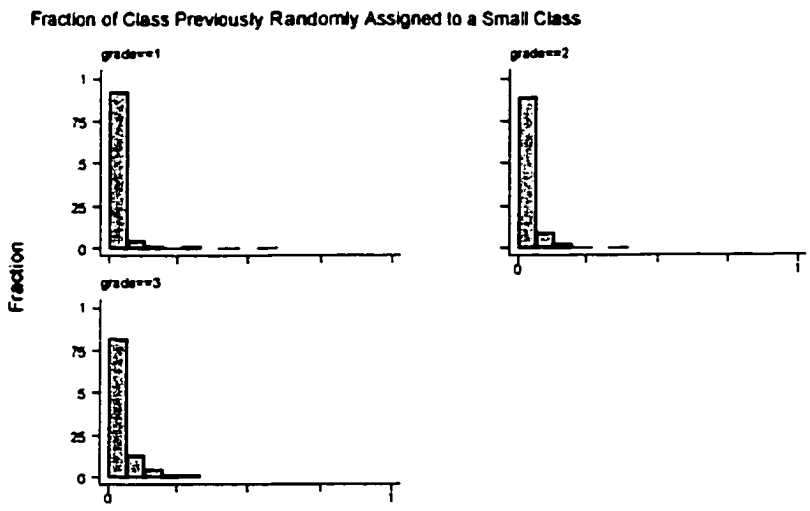


Figure 2: Class Level Histograms

**Figure 3: Between Class Partial Plot**
(Net of Class Type and Other Covariates)

**Table 1**
**Mean Characteristics of Switchers, Stayers, and New Entrants, Conditional on School Effects**

| | First Grade | Second Grade | Third Grade |
|---|---|---|---|
| **White** | **.67** | **.65** | **.67** |
| Switch to Small Class | .67 | .63 | .65 |
| | [248] | [192] | [207] |
| Switch to Regular Class | .74 | .62 | .67 |
| | [108] | [47] | [72] |
| Stay in Small Class | .68 | .66 | .67 |
| | [1293] | [1435] | [1564] |
| Stay in Regular Class | .68 | .65 | .67 |
| | [2867] | [3375] | [3570] |
| New Entrant, Small Class | .63 | .63 | .68 |
| | [380] | [339] | [368] |
| New Entrant, Regular Class | .65 | .65 | .67 |
| | [1904] | [1246] | [894] |
| **Girl** | **.48** | **.48** | **.48** |
| Switch to Small Class | .48 | .50 | .50 |
| | [248] | [192] | [207] |
| Switch to Regular Class | .53 | .52 | .55 |
| | [108] | [47] | [72] |
| Stay in Small Class | .49 | .50 | .50 |
| | [1293] | [1435] | [1564] |
| Stay in Regular Class | .50 | .49 | .48 |
| | [2867] | [3375] | [3571] |
| New Entrant, Small Class | .48 | .42 | .43 |
| | [383] | [366] | [373] |
| New Entrant, Regular Class | .45 | .45 | .47 |
| | [1917] | [1306] | [908] |
| **Free Lunch (status in previous grade)** | **.52** | **.51** | **.51** |
| Switch to Small Class | .47 | .51 | .47 |
| | [246] | [192] | [202] |
| Switch to Regular Class | .48 | .48 | .52 |
| | [107] | [45] | [71] |
| Stay in Small Class | .44 | .44 | .44 |
| | [1288] | [1408] | [1509] |
| Stay in Regular Class | .45 | .48 | .48 |
| | [2858] | [3284] | [3410] |
| New Entrant, Small Class (status in current grade) | .69 | .76 | .74 |
| | [372] | [357] | [356] |

| | | | |
|---|---|---|---|
| New Entrant, Regular Class | .71 | .72 | .77 |
| (status in current grade) | [1867] | [1235] | [844] |
| **Percentile Test Score** | **50.79** | **50.61** | **51.02** |
| **(In previous grade)** | | | |
| Switch to Small Class | 51.18 | 52.61 | 51.42 |
| | [230] | [188] | [195] |
| Switch to Regular Class | 51.63 | 57.59 | 51.06 |
| | [101] | [45] | [62] |
| Stay in Small Class | 57.92 | 60.03 | 56.96 |
| | [1212] | [1418] | [1473] |
| Stay in Regular Class | 52.26 | 53.36 | 51.17 |
| | [2706] | [3330] | [3339] |
| New Entrant, Small Class | 42.95 | 43.71 | 40.31 |
| (score in current grade) | [357] | [255] | [276] |
| New Entrant, Regular Class | 39.65 | 39.93 | 38.05 |
| (score in current grade) | [1823] | [1017] | [750] |

Notes: Sample sizes of the relevant groups are in brackets. Regular size classes and regular/aide classes have been collapsed into one group called "regular". The sample sizes don't match up within grades across variables due to missing observations. For the time-varying characteristics (free lunch and percentile test score), the switchers' and stayers' means are computed based on the *previous* grade, while the new entrants' means are based on the *current* grade.

**Table 2**
**Composition of Class Types in Each Grade**
**Number of Students Broken-Out by Random Assignment Status**

| | Small | Regular | Total |
|---|---|---|---|
| Kindergarten | | | |
| Randomly Assigned | 1900 | 4425 | 6325 |
| Total | 1900 | 4425 | 6325 |
| First grade | | | |
| Previously Randomly Assigned | 1293 | 2867 | 4160 |
| New Entrants | 384 | 1929 | 2313 |
| Switchers | 248 | 108 | 356 |
| (from previous year) | (248) | (108) | (356) |
| Total | 1925 | 4904 | 6829 |
| Second grade | | | |
| Previously Randomly Assigned | 1273 | 3402 | 4675 |
| New Entrants | 366 | 1313 | 1679 |
| Switchers | 377 | 109 | 486 |
| (from previous year) | (192) | (47) | (239) |
| Total | 2016 | 4824 | 6840 |
| Third Grade | | | |
| Previously Randomly Assigned | 1276 | 3567 | 4843 |
| New Entrants | 373 | 908 | 1281 |
| Switchers | 525 | 153 | 678 |
| (from previous year) | (207) | (72) | (279) |
| Total | 2174 | 4628 | 6802 |

Notes: Regular and regular/aide students are grouped together. "Previously randomly assigned" refers to students having been randomly assigned in an earlier grade to the class type column under consideration, e.g. in the column for small classes, the previously randomly assigned students were randomly assigned to a *small* class in their grade of entry. "Switchers" refers to students who were not in the class type, in the relevant grade, to which they were randomly assigned. In parentheses under the switchers' rows are the number of students who switched class type from the previous year. The "total" column sums horizontally across the small and regular class columns. The "total" rows sum vertically across rows within each grade, not including numbers in parentheses.

**Table 3**
**OLS Estimates of the Experimental Effect on Individual Test Scores by Grade**

| | Kindergarten | First Grade | Second Grade | Third Grade |
|---|---|---|---|---|
| Small class | 5.13 | 7.31 | 5.94 | 4.76 |
| | (1.25) | (1.17) | (1.27) | (1.26) |
| Regular/aide class | .22 | 1.57 | 1.64 | -.51 |
| | (1.14) | (.97) | (1.07) | (1.16) |
| White | 9.38 | 8.39 | 8.00 | 7.15 |
| | (1.38) | (1.19) | (1.25) | (1.45) |
| Girl | 4.46 | 3.17 | 3.34 | 3.21 |
| | (.63) | (.57) | (.59) | (.68) |
| Free lunch | -13.03 | -13.02 | -13.24 | -12.21 |
| | (.79) | (.87) | (.72) | (.82) |
| White teacher | -1.02 | -4.13 | 1.08 | 1.23 |
| | (2.20) | (1.98) | (1.79) | (1.79) |
| Master's degree | .76 | .34 | -.65 | 1.67 |
| | (1.13) | (1.08) | (1.12) | (1.22) |
| Teacher's experience | .26 | .04 | .07 | .05 |
| | (.11) | (.06) | (.07) | (.06) |
| School fixed effects | Yes | Yes | Yes | Yes |
| $R^2$ | .32 | .31 | .30 | .24 |
| Number of obs | 5701 | 6437 | 5747 | 5816 |

Notes: Robust standard errors that allow for a correlation of the residuals among members of the same class are in parentheses. A constant is included in all regressions.

**Table 4**
**OLS Estimates of Class Size and Peer Group Effects by Grade: Dependent Variable is Individual Test Score**

| | First Grade | Second Grade | Third Grade |
|---|---|---|---|
| Peers' Mean Test Score | .58 | .58 | .57 |
| | (.04) | (.04) | (.04) |
| Small class | 2.66 | 2.18 | 1.67 |
| | (.58) | (.53) | (.58) |
| Regular/aide class | .54 | .49 | -.30 |
| | (.43) | (.45) | (.51) |
| White | 8.51 | 8.09 | 7.08 |
| | (1.17) | (1.24) | (1.43) |
| Girl | 3.14 | 3.27 | 3.41 |
| | (.57) | (.60) | (.68) |
| Free lunch | -12.97 | -12.95 | -12.28 |
| | (.86) | (.70) | (.81) |
| White teacher | -2.11 | .53 | .14 |
| | (.82) | (.74) | (.78) |
| Master's degree | .26 | .03 | .61 |
| | (.48) | (.47) | (.52) |
| Teacher's experience | .02 | .03 | .02 |
| | (.03) | (.03) | (.03) |
| School fixed effects | Yes | Yes | Yes |
| Number of obs | 6437 | 5747 | 5816 |
| Normalized Peer Effect | 4.00 | 3.08 | 3.28 |
| | (.28) | (.21) | (.23) |

Notes: Robust standard errors that allow for a correlation of the residuals among members of the same class are in parentheses. A constant is included in all regressions. The normalized peer effect is constructed by considering the thought experiment of moving a student from a regular size class to a small class allowing the quality of the student's peers to change, yet holding class size constant. Formally, it is computed by multiplying the coefficient on peer's mean test score by the difference in mean peers' test scores for small and regular classes. For example, in third grade, moving from a regular class to a small class entails an increase in mean peers' score from 49.14 to 54.90, yielding a normalized peer effect of .57*(54.90 - 49.14) = 3.28 percentile points. This normalized peer effect can be compared directly with the small class coefficient to shed some light on the relative magnitudes of each.

**Table 5**
**First Stage of Instrumental Variables Estimation:**
**Dependent Variable is Peers' Mean Test Score**

| | First Grade | Second Grade | Third Grade |
|---|---|---|---|
| Fraction of Peers Randomly Assigned to a Small Class in Kindergarten | 2.37<br>(3.56) | 6.85<br>(3.84) | 17.37<br>(3.81) |
| Fraction of Peers Randomly Assigned to a Small Class in First Grade | ------ | 4.46<br>(8.20) | 3.20<br>(9.10) |
| Fraction of Peers Randomly Assigned to a Small Class in Second Grade | ------ | ------ | -4.11<br>(8.00) |
| Small class | 6.39<br>(2.50) | 2.44<br>(2.57) | -1.53<br>(2.23) |
| Regular/aide class | 1.79<br>(.99) | 1.91<br>(1.09) | -.50<br>(1.14) |
| White teacher | -3.30<br>(2.04) | 1.00<br>(1.84) | 1.05<br>(1.84) |
| Master's degree | .14<br>(1.09) | -1.35<br>(1.18) | 1.86<br>(1.19) |
| Teacher's experience | .04<br>(.06) | .05<br>(.07) | .05<br>(.06) |
| F-statistic for Joint Test of Peer Variables (p-value) | 0.44<br>(.509) | 1.64<br>(.197) | 7.65<br>(.0001) |
| School fixed effects | Yes | Yes | Yes |
| $R^2$ | .73 | .70 | .67 |
| Number of obs | 6437 | 5747 | 5816 |

Notes: Robust standard errors that allow for a correlation of the residuals among members of the same class are in parentheses. A constant is included in all regressions, as are student characteristics (white, girl, free lunch).

**Table 6**
**Instrumental Variables Estimates of Class Size and Peer Group Effects by Grade: Peers' Mean Test Score Instrumented by Random Assignment Status of Peers**

| | First Grade | Second Grade | Third Grade |
|---|---|---|---|
| Peers' Mean Test Score | .30 | .86 | .92 |
| | (1.00) | (.12) | (.04) |
| Small class | 4.91 | .38 | -.17 |
| | (7.94) | (.78) | (.32) |
| Regular/aide class | 1.04 | -.05 | -.17 |
| | (1.92) | (.30) | (.19) |
| White | 8.45 | 8.13 | 7.04 |
| | (1.19) | (1.25) | (1.44) |
| Girl | 3.16 | 3.23 | 3.53 |
| | (.57) | (.60) | (.69) |
| Free lunch | -12.99 | -12.81 | -12.32 |
| | (.87) | (.71) | (.82) |
| White teacher | -3.07 | .26 | -.51 |
| | (3.69) | (.30) | (.28) |
| Master's degree | .30 | .36 | -.02 |
| | (.78) | (.27) | (.20) |
| Teacher's experience | .03 | .02 | -.003 |
| | (.06) | (.01) | (.01) |
| School fixed effects | Yes | Yes | Yes |
| Number of obs | 6437 | 5747 | 5816 |
| Normalized Peer Effect | 2.05 | 4.49 | 4.66 |
| | (6.77) | (.63) | (.20) |

Notes: Robust standard errors that allow for a correlation of the residuals among members of the same class are in parentheses. A constant is included in all regressions. The normalized peer effect is constructed by considering the thought experiment of moving a student from a regular size class to a small class allowing the quality of the student's peers to change, yet holding class size constant. Formally, it is computed by multiplying the coefficient on peer's mean test score by the difference in mean predicted peers' test scores for small and regular classes. For example, in third grade, moving from a regular class to a small class entails an increase in mean predicted peers' score from 49.44 to 54.51, yielding a normalized peer effect of .92*(54.51 - 49.44) = 4.66 percentile points. This normalized peer effect can be compared directly with the small class coefficient to shed some light on the relative magnitudes of each.

**Table 7**
**Instrumental Variables Estimates of Peer Group Effects by Grade: Looking Within Class Type by Instrument Sets**

| | First Grade | Second Grade | Third Grade |
|---|---|---|---|
| Instruments Are Percent of Peers Randomly Assigned To a Small Class: | | | |
| Both Class Types | .30<br>(1.00) | .86<br>(.12) | .92<br>(.04) |
| Small Classes | 1.72<br>(.71) | 1.01<br>(.10) | .89<br>(.05) |
| Regular Classes | .86<br>(.12) | -.56<br>(4.66) | 1.00<br>(.09) |
| Instruments are Percent of Peers Entering in Each Grade: | | | |
| Both Class Types | .61<br>(.16) | .68<br>(.08) | .72<br>(.05) |
| Small Classes | -.81<br>(2.05) | .60<br>(.13) | .65<br>(.10) |
| Regular Classes | .52<br>(.27) | .43<br>(.21) | .39<br>(.22) |
| Instruments are Percent of Peers Switching in Each Grade: | | | |
| Small Classes | .93<br>(.13) | .81<br>(.10) | 1.06<br>(.09) |
| Regular Classes | .86<br>(.12) | -.56<br>(4.66) | 1.00<br>(.09) |

Notes: Each cell represents a separate regression. Robust standard errors that allow for a correlation among members of the same class are in parentheses. A constant is included in all regressions, as are student characteristics, teacher characteristics, and school fixed effects.

**Table 8**
**Non-Linearities in Peer Group Effects, Class Level Estimates:**
**Dependent Variable is Class Mean Test Score**

| | First Grade | Second Grade | Third Grade |
|---|---|---|---|
| Percent of Kids Randomly Assigned to a Small Class is Between 0 and 20% | ------ [214] | ----- [207] | ----- [200] |
| Percent of Kids Randomly Assigned to a Small Class is Between 20% and 40% | .57 (4.22) [15] | 1.14 (4.60) [10] | 1.49 (3.62) [14] |
| Percent of Kids Randomly Assigned to a Small Class is Between 40% and 60% | 5.37 (4.24) [24] | 5.64 (4.07) [34] | 2.19 (3.20) [39] |
| Percent of Kids Randomly Assigned to a Small Class is Between 60% and 80% | 3.85 (4.12) [45] | 4.75 (4.06) [39] | 8.72 (3.14) [43] |
| Percent of Kids Randomly Assigned to a Small Class is Between 80% and 100% | 2.23 (4.23) [40] | 5.07 (4.04) [40] | 10.98 (3.30) [33] |
| Number of obs | 338 | 330 | 329 |

Notes: Standard errors are in parentheses. Sample size of each group is in brackets. Additional covariates in each regression are a constant, class type, white teacher, teacher has a masters, teacher's experience, and school dummies.

**Table 9**
**Non-Linearities in Peer Group Effects, Class Level Estimates Including Constancy of Classmates:**
**Dependent Variable is Class Mean Test Score**

| | First Grade | Second Grade | Third Grade |
|---|---|---|---|
| Percent of Kids Randomly Assigned to a Small Class is Between 0% and 20% | ----- [214] | ----- [207] | ----- [200] |
| Percent of Kids Randomly Assigned to a Small Class is Between 20% and 40% | 1.01 (4.24) [15] | 2.13 (4.59) [10] | 1.88 (3.63) [14] |
| Percent of Kids Randomly Assigned to a Small Class is Between 40% and 60% | 5.98 (4.24) [24] | 3.69 (4.11) [34] | 2.11 (3.22) [39] |
| Percent of Kids Randomly Assigned to a Small Class is Between 60% and 80% | 4.79 (4.18) [45] | 2.83 (4.07) [39] | 8.48 (3.16) [43] |
| Percent of Kids Randomly Assigned to a Small Class is Between 80% and 100% | 5.64 (4.72) [40] | 1.87 (4.13) [40] | 9.84 (3.37) [33] |
| Average Fraction of Class Previously Together is Between 0% and 20% | ----- [207] | ----- [48] | ----- [18] |
| Average Fraction of Class Previously Together is Between 20% and 40% | -.81 (1.90) [97] | 3.80 (2.23) [106] | 1.80 (2.96) [84] |
| Average Fraction of Class Previously Together is Between 40% and 60% | -.93 (3.25) [25] | 6.49 (2.52) [85] | 4.31 (3.16) [98] |
| Average Fraction of Class Previously Together is Between 60% and 80% | -11.42 (6.24) [5] | 8.71 (2.87) [53] | 4.00 (3.40) [88] |

| | | | |
|---|---|---|---|
| Average Fraction of Class | -11.28 | 7.09 | 6.12 |
| Previously Together is | (7.03) | (3.11) | (3.64) |
| Between 80% and 100% | [4] | [38] | [41] |
| Number of obs | 338 | 330 | 329 |

Notes: Standard errors are in parentheses. Sample size of each group is in brackets. Additional covariates in each regression are the same as in Table 8: class type, white teacher, teacher has a masters, teacher's experience, and school fixed effects.

**Table 10**
**Between and Within Class Estimates:**
**Dependent Variable is Class Mean (or Individual) Test Score**

| | First Grade | Second Grade | Third Grade |
|---|---|---|---|
| **Between Class Estimates:** | | | |
| Fraction of Class Previously Randomly Assigned to a Small Class | 1.53<br>(4.13) | 4.26<br>(4.17) | 13.77<br>(3.73) |
| Small | 6.48<br>(3.01) | 2.87<br>(2.97) | -3.74<br>(2.63) |
| Regular/aide Class | 1.71<br>(1.33) | 1.28<br>(1.43) | -1.07<br>(1.49) |
| Fraction White | 10.00<br>(10.31) | 15.47<br>(11.09) | 12.44<br>(11.18) |
| Fraction Girl | 7.08<br>(7.39) | 10.09<br>(7.67) | .96<br>(6.88) |
| Fraction Free lunch | -12.90<br>(4.94) | -24.08<br>(5.85) | -16.96<br>(6.04) |
| Number of obs | 336 | 320 | 322 |
| **Within Class Estimates:** | | | |
| Individual Previously Randomly Assigned to a Small Class | 3.64<br>(1.09) | 1.53<br>(1.14) | 2.33<br>(1.08) |
| White | 8.25<br>(1.06) | 7.81<br>(1.15) | 6.84<br>(1.26) |
| Girl | 3.06<br>(.54) | 2.97<br>(.57) | 3.42<br>(.60) |
| Free lunch | -12.88<br>(.66) | -12.75<br>(.71) | -12.18<br>(.74) |
| Number of obs | 6449 | 5829 | 5878 |
| Implied Peer Coefficient | -1.38 | .64 | .83 |

Notes: Standard errors are in parentheses. A constant and school fixed effects are included in all regressions. Teacher characteristics are included in the between class regressions. The implied peer coefficient is calculated as 1 - (within coefficient)/(between coefficient).

**Appendix Table 1**
**Class Level Reduced Form Estimates Including Fraction of Class Entering in Each Grade:**
**Dependent Variable is Class Mean Test Score**

| | First Grade | Second Grade | Third Grade |
|---|---|---|---|
| Fraction of Kids Randomly Assigned to a Small Class in Kindergarten | -2.21<br>(4.45) | -2.19<br>(4.69) | 10.31<br>(4.34) |
| Fraction of Kids Randomly Assigned to a Small Class in First Grade | ----- | 9.87<br>(10.83) | 5.89<br>(11.29) |
| Fraction of Kids Randomly Assigned to a Small Class in Second Grade | ----- | ----- | 6.14<br>(10.22) |
| Small class | 7.12<br>(3.01) | 2.30<br>(3.03) | -3.82<br>(2.63) |
| Regular/aide class | 1.55<br>(1.32) | 1.32<br>(1.41) | -1.50<br>(1.40) |
| Fraction of Class Entering in First Grade | -11.87<br>(5.22) | -26.29<br>(7.87) | -19.15<br>(8.13) |
| Fraction of Class Entering in Second Grade | ----- | -17.53<br>(5.02) | -21.21<br>(7.07) |
| Fraction of Class Entering in Third Grade | ----- | ----- | -28.89<br>(6.16) |
| School fixed effects | Yes | Yes | Yes |
| $R^2$ | .73 | .70 | .70 |
| F-statistic for Joint Test of Peer Variables<br>(p-value) | 0.25<br>(.620) | 0.57<br>(.567) | 2.19<br>(.090) |
| Number of obs | 338 | 330 | 329 |

Notes: Standard errors are in parentheses. A constant is included in all regressions. Additional covariates include teacher characteristics.

**Appendix Table 2**
**Instrumental Variables Estimates of Class Size and Peer Group Effects by Grade: Peers' Mean Test Score Instrumented by Random Assignment Status of Peers, Individual PRASC Included as a Covariate**

| | First Grade | Second Grade | Third Grade |
|---|---|---|---|
| Peers' Mean Test Score | -.23 | .70 | .83 |
| | (1.67) | (.22) | (.08) |
| Previously Randomly Assigned to a Small Class | 1.12 | -.37 | .81 |
| | (1.14) | (1.04) | (1.11) |
| Small Class Currently | 7.54 | .44 | -1.34 |
| | (13.36) | (1.15) | (.48) |
| Regular/aide class | 1.98 | .13 | -.46 |
| | (3.21) | (.52) | (.26) |
| Attended Kindergarten (In a STAR school) | 4.47 | 6.00 | 6.63 |
| | (.93) | (.72) | (.69) |
| White | 7.97 | 7.60 | 6.77 |
| | (1.23) | (1.23) | (1.43) |
| Girl | 3.00 | 2.92 | 3.07 |
| | (.57) | (.61) | (.68) |
| Free lunch | -12.43 | -11.85 | -10.90 |
| | (.89) | (.70) | (.84) |
| White teacher | -4.95 | .58 | -.45 |
| | (6.21) | (.51) | (.42) |
| Master's degree | .43 | .24 | .14 |
| | (1.34) | (.44) | (.30) |
| Teacher's experience | .05 | .03 | .01 |
| | (.10) | (.02) | (.01) |
| School fixed effects | Yes | Yes | Yes |
| Number of obs | 6437 | 5747 | 5816 |
| Normalized Peer Effect | -1.56 | 3.57 | 4.11 |
| | (11.33) | (1.12) | (.40) |

Notes: Robust standard errors that allow for a correlation of the residuals among members of the same class are in parentheses. A constant is included in all regressions.

**Appendix Table 3**
**Individual Level Reduced Form:**
**Dependent Variable is Individual Test Score**

| | First Grade | Second Grade | Third Grade |
|---|---|---|---|
| Individual Randomly Assigned to a Small Class in Kindergarten | 3.68<br>(1.01) | 2.91<br>(1.12) | 4.23<br>(1.15) |
| Individual Randomly Assigned to a Small Class in First Grade | ------ | -4.07<br>(1.73) | -.94<br>(2.13) |
| Individual Randomly Assigned to a Small Class in Second Grade | ------ | ------ | .03<br>(1.68) |
| Fraction of Peers Randomly Assigned to a Small Class in Kindergarten | -1.80<br>(3.46) | 3.74<br>(3.79) | 13.05<br>(3.75) |
| Fraction of Peers Randomly Assigned to a Small Class in First Grade | ----- | 6.86<br>(8.05) | 4.25<br>(9.04) |
| Fraction of Peers Randomly Assigned to a Small Class in Second Grade | ----- | ----- | -1.85<br>(7.49) |
| Small class | 6.03<br>(2.38) | 2.07<br>(2.57) | -2.42<br>(2.15) |
| Regular/aide class | 1.59<br>(.97) | 1.58<br>(1.06) | -.64<br>(1.14) |
| F-statistic for Joint Test of Peer Variables (p-value) | 0.27<br>(.604) | 0.75<br>(.472) | 7.65<br>(.0001) |
| Number of obs | 6437 | 5747 | 5816 |

Notes: Robust standard errors that allow for a correlation of the residuals among members of the same class are in parentheses. A constant is included in all regressions, as are student characteristics, teacher characteristics, and school fixed effects.