Boozer, Michael A.; Cacciola, Stephen E.

Working Paper

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

Center Discussion Paper, No. 832

Provided in Cooperation with:
Yale University, Economic Growth Center (EGC)

Suggested Citation: Boozer, Michael A.; Cacciola, Stephen E. (2001) : Inside the 'Black Box' of Project STAR: Estimation of Peer Effects Using Experimental Data, Center Discussion Paper, No. 832, Yale University, Economic Growth Center, New Haven, CT

This Version is available at:
https://hdl.handle.net/10419/98316

ECONOMIC GROWTH CENTER

YALE UNIVERSITY

P.O. Box 208269
New Haven, CT 06520-8269

CENTER DISCUSSION PAPER NO. 832

INSIDE THE 'BLACK BOX' OF PROJECT STAR: ESTIMATION OF PEER EFFECTS USING EXPERIMENTAL DATA[1]

Michael A. Boozer
Yale University

and

Stephen E. Cacciola
Yale University

June 2001

Note: Center Discussion Papers are preliminary materials circulated to stimulate discussions and critical comments.

[1]Initial notes (Section Five): May 1997.

We thank Dean Hyslop, Ann Stevens, Jenny Hunt, Paul Schultz, Andrew Foster, and seminar audiences at Yale, Brown, and CUNY for helpful comments. The identification strategy in this paper was inspired by George Akerlof's (1997) recounting of Eugene Lang's scholarship intervention in Harlem.

Inside the 'Black Box' of Project Star: Estimation of Peer Effects Using Experimental Data

Michael A. Boozer
michael.boozer@yale.edu

and

Stephen E. Cacciola
stephen.cacciola@yale.edu

Abstract

The credible identification of endogenous peer group effects—i.e. social multiplier or feedback effects—has long eluded social scientists. We argue that such effects are most credibly identified by a randomly assigned social program which operates at differing intensities within and between peer groups. The data we use are from Project STAR, a class size reduction experiment conducted in Tennessee elementary schools. In these data, classes were comprised of varying fractions of students who had previously been exposed to the Small class treatment, creating class groupings of varying experimentally induced quality. We use this variation in class group quality to estimate the spillover effect. We find that when allowance is made for this 'feedback' effect of prior exposure to the Small class treatment, 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.

JEL Classification: Z13, C51, C81, I21, C23

Keywords: Peer Effects; Data with a Group Structure; Organization of Schooling; Experimental Evidence

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.[4] 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

[4]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.

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 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.[5] 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 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

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

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.[6] 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 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[7], 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

[6]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.

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

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 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[8]; 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 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

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

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.[9] 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.[10]

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:

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

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.[11] The key regressor of interest is the sample mean of

[9] 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.

[10] 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.

[11] 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 as-

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}.

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 β 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

sumptions 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*.

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.[12] 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 composition, 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

[12]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.

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.[13] 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 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*[14] 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. Further-

[13] 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.

[14] 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.

more, 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, 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.[15] 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

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

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 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.[16] 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

[16] 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.

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 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.[17] The remarkable stability of the es-

[17] 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

timated coefficients across grades on the 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.[18]

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 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.

---

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 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.

[18] 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.

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.[19]

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.

[19] 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 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.

Notice also that because we are instrumenting for a grouped version of the dependent variable, the first-stage regression is also almost the reduced form for the two equation system at the individual level.[20] 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 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.

[20]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.

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 β 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.[21] 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 brevity. Unfortunately, as the relevant variation here occurs at the class level and we have

[21] 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.

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 regresssor 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).[22] 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 estimate of 1 regardless of the underlying data

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

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.[23]

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

[23] 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.

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 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 mechanisms, 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.[24] 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 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

[24] 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 $\bar{d}_{-i,j}$.

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:

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

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 β 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 β 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:

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

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:

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

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 β 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 kth element of the coefficient vectors on the xs 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:

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

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 β 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 β (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.[25]

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 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

[25] 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).

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.[26] The 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:

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

[26] 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.

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 β 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:

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

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:

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

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 ρ. 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), 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,' Working Paper, Economic Growth Center, 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), 557-577.

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

[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), 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, Cambridge.

[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), 487-535.

[11] Krueger, Alan B., (1999), 'Experimental Estimates of Education Production Functions,' *Quarterly Journal of Economics*, 114(2), 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), 1-28.

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

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

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

[16] Manski, Charles F., (2000), 'Economic Analysis of Social Interactions,' *The Journal of Economic Perspectives*, 14(3), 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: Cambridge.

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

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

[20] Senesky, Sarah E., (2000), 'Commuting Time as a Measure of Employment Costs,' Working Paper, 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, Nashville.

[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, Williamstown, Mass.

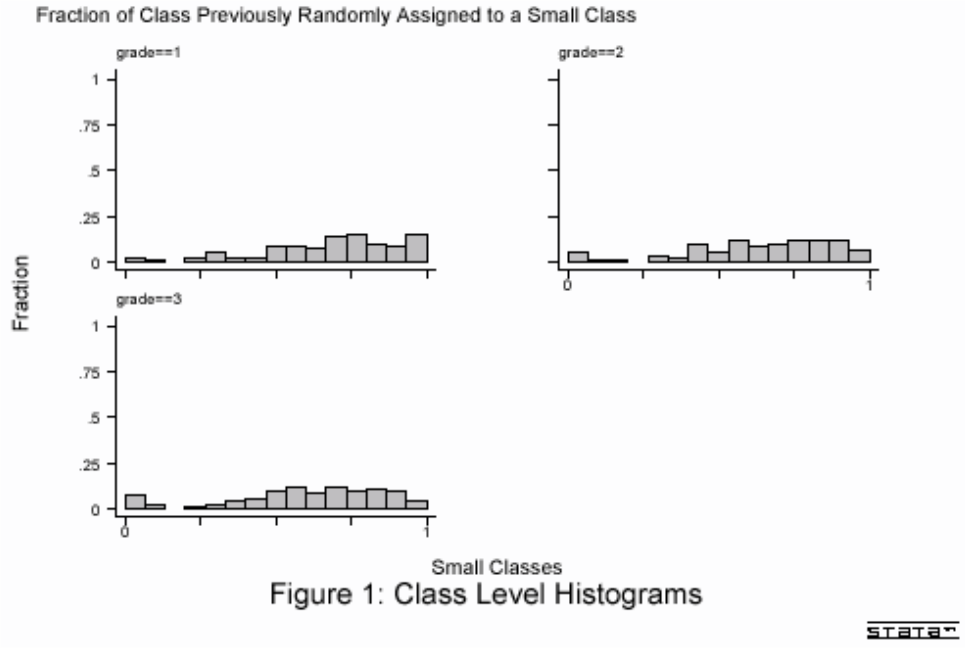

Figure 1: Class Level Histograms

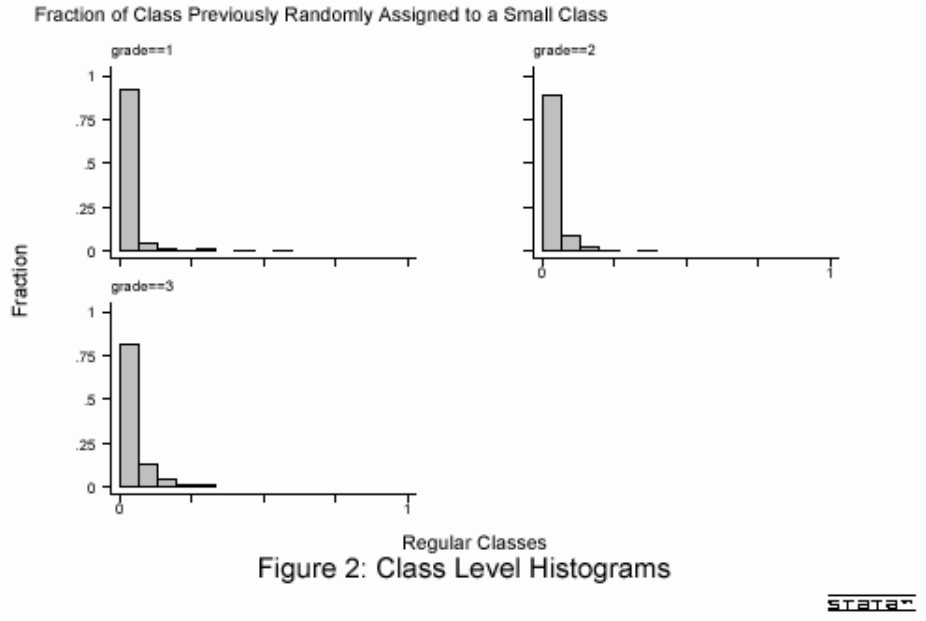

Figure 2: Class Level Histograms

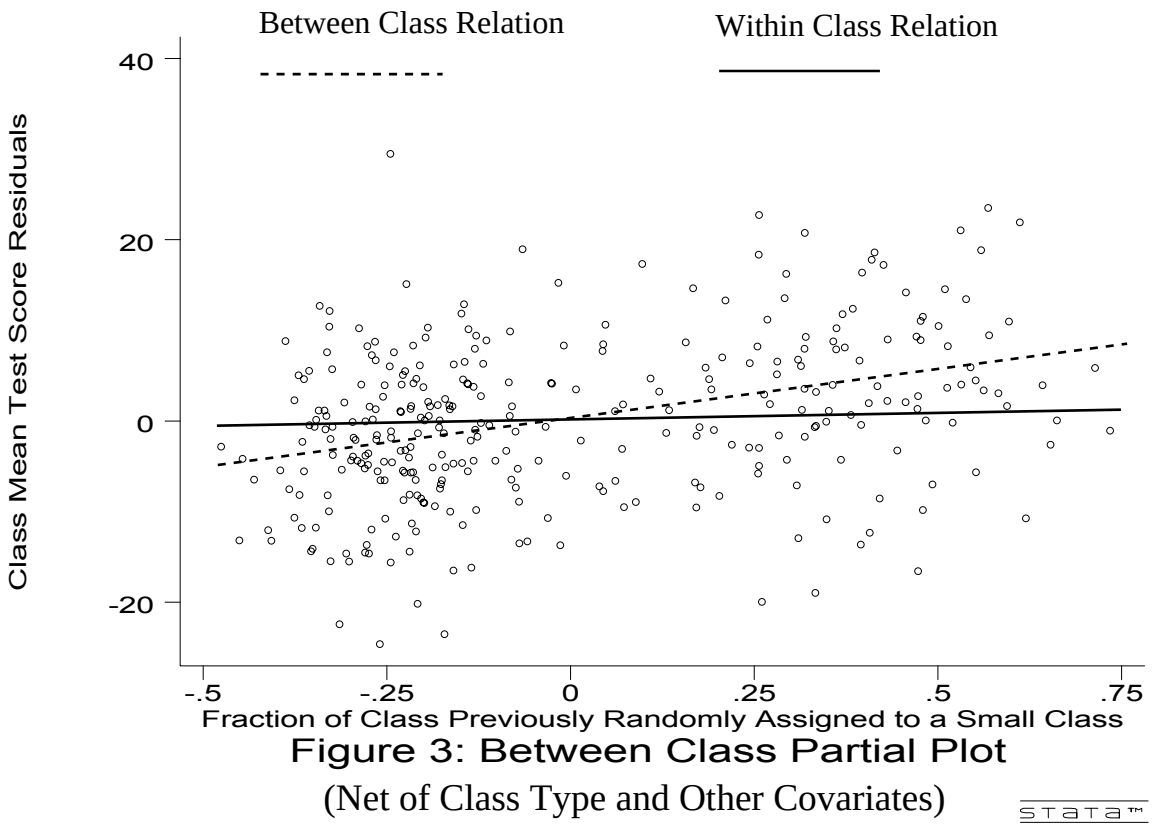

Figure 3: Between Class Partial Plot

(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	.69	.76	.74
(status in current grade)	[372]	[357]	[356]
New Entrant, Regular Class	.71	.72	.77
(status in current grade)	[1867]	[1235]	[844]

Percentile Test Score (In previous grade)	**50.79**	**50.61**	**51.02**
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 (3.56)	6.85 (3.84)	17.37 (3.81)
Fraction of Peers Randomly Assigned to a Small Class in First Grade	------	4.46 (8.20)	3.20 (9.10)
Fraction of Peers Randomly Assigned to a Small Class in Second Grade	------	------	-4.11 (8.00)
Small class	6.39 (2.50)	2.44 (2.57)	-1.53 (2.23)
Regular/aide class	1.79 (.99)	1.91 (1.09)	-.50 (1.14)
White teacher	-3.30 (2.04)	1.00 (1.84)	1.05 (1.84)
Master's degree	.14 (1.09)	-1.35 (1.18)	1.86 (1.19)
Teacher's experience	.04 (.06)	.05 (.07)	.05 (.06)
F-statistic for Joint Test of Peer Variables (p-value)	0.44 (.509)	1.64 (.197)	7.65 (.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 (1.00)	.86 (.12)	.92 (.04)
Small Classes	1.72 (.71)	1.01 (.10)	.89 (.05)
Regular Classes	.86 (.12)	-.56 (4.66)	1.00 (.09)
Instruments are Percent of Peers Entering in Each Grade:			
Both Class Types	.61 (.16)	.68 (.08)	.72 (.05)
Small Classes	-.81 (2.05)	.60 (.13)	.65 (.10)
Regular Classes	.52 (.27)	.43 (.21)	.39 (.22)
Instruments are Percent of Peers Switching in Each Grade:			
Small Classes	.93 (.13)	.81 (.10)	1.06 (.09)
Regular Classes	.86 (.12)	-.56 (4.66)	1.00 (.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 (4.13)	4.26 (4.17)	13.77 (3.73)
Small	6.48 (3.01)	2.87 (2.97)	-3.74 (2.63)
Regular/aide Class	1.71 (1.33)	1.28 (1.43)	-1.07 (1.49)
Fraction White	10.00 (10.31)	15.47 (11.09)	12.44 (11.18)
Fraction Girl	7.08 (7.39)	10.09 (7.67)	.96 (6.88)
Fraction Free lunch	-12.90 (4.94)	-24.08 (5.85)	-16.96 (6.04)
Number of obs	336	320	322
Within Class Estimates:			
Individual Previously Randomly Assigned to a Small Class	3.64 (1.09)	1.53 (1.14)	2.33 (1.08)
White	8.25 (1.06)	7.81 (1.15)	6.84 (1.26)
Girl	3.06 (.54)	2.97 (.57)	3.42 (.60)
Free lunch	-12.88 (.66)	-12.75 (.71)	-12.18 (.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 (4.45)	-2.19 (4.69)	10.31 (4.34)
Fraction of Kids Randomly Assigned to a Small Class in First Grade	-----	9.87 (10.83)	5.89 (11.29)
Fraction of Kids Randomly Assigned to a Small Class in Second Grade	-----	-----	6.14 (10.22)
Small class	7.12 (3.01)	2.30 (3.03)	-3.82 (2.63)
Regular/aide class	1.55 (1.32)	1.32 (1.41)	-1.50 (1.40)
Fraction of Class Entering in First Grade	-11.87 (5.22)	-26.29 (7.87)	-19.15 (8.13)
Fraction of Class Entering in Second Grade	-----	-17.53 (5.02)	-21.21 (7.07)
Fraction of Class Entering in Third Grade	-----	-----	-28.89 (6.16)
School fixed effects	Yes	Yes	Yes
R^2	.73	.70	.70
F-statistic for Joint Test of Peer Variables (p-value)	0.25 (.620)	0.57 (.567)	2.19 (.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	1.12	-.37	.81
Assigned to a Small Class	(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	4.47	6.00	6.63
(In a STAR school)	(.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 (1.01)	2.91 (1.12)	4.23 (1.15)
Individual Randomly Assigned to a Small Class in First Grade	------	-4.07 (1.73)	-.94 (2.13)
Individual Randomly Assigned to a Small Class in Second Grade	------	------	.03 (1.68)
Fraction of Peers Randomly Assigned to a Small Class in Kindergarten	-1.80 (3.46)	3.74 (3.79)	13.05 (3.75)
Fraction of Peers Randomly Assigned to a Small Class in First Grade	-----	6.86 (8.05)	4.25 (9.04)
Fraction of Peers Randomly Assigned to a Small Class in Second Grade	-----	-----	-1.85 (7.49)
Small class	6.03 (2.38)	2.07 (2.57)	-2.42 (2.15)
Regular/aide class	1.59 (.97)	1.58 (1.06)	-.64 (1.14)
F-statistic for Joint Test of Peer Variables (p-value)	0.27 (.604)	0.75 (.472)	7.65 (.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.