Identifying Social Interactions through Excess Variance Contrasts

by

Bryan S. Graham*

(Initial Draft: June 2003)

(This Draft: May 5th, 2005)

Abstract

This paper outlines a new method for detecting and assessing the strength of social interactions based on contrasts in excess variance across social groups of exogenously differing sizes. An attractive feature of the approach is its robustness to the presence of group-level heterogeneity and sorting. The proposed estimation strategy is used to test for the presence of peer effects in learning using data from the Tennessee class size reduction experiment Project STAR. Size-induced contrasts of excess variance provide a powerful mechanism for detecting peer group effects in this dataset. Switching from classroom where mean peer ability is at the 25th percentile of the ability distribution to one where it is at the 75th percentile is associated with changes in math and reading achievement scores of 0.9 and 1.1 standard deviations respectively. These estimates suggest that, at minimum, differences in peer composition are at least as important as those in teacher quality for explaining variation in academic achievement within Project STAR schools. While tests based on excess variance contrasts provide strong evidence of peer group effects, conventional regression-based excess sensitivity tests do not. Calibrating asymptotic power functions for the two tests to the Project STAR data suggests that across repeated samples the odds of detecting social interactions are roughly 20 to 30 times greater with the proposed excess variance test. Generalized method of moments provides a unified framework for estimation and inference. The proposed approach is straightforward to implement using standard software.

JEL Classification: C31, C39, I29, J24.

Key Words: Social Interactions, Peer Group Effects, Reflection Problem, Linear-in-Means, Project STAR, Educational Production, Covariance Models.

*I would like to thank Gary Chamberlain, Larry Katz, Michael Kremer, and Caroline Hoxby for their encouragement as well as many helpful comments, corrections and suggestions. Useful feedback from William Brock, David Card, Steven Durlauf, Bo Honoré, Guido Imbens, Richard Murnane, Mark Watson, John Willett, Jeff Zabel and participants in the Banff Workshop on Social Interactions as well as seminars at Harvard, Harvard-MIT, UCSD, Dartmouth, Princeton, Stanford, Stanford-GSB, UT-Austin, Berkeley, Chicago, Wisconsin, LSE, Yale, IFS, Toulouse-GREMAQ and Munich is also gratefully acknowledged. Financial support provided by a National Science Foundation Graduate Fellowship, the Program on Justice, Welfare and Economics at Harvard University and by the Program of Fellowships for Junior Scholars, the MacArthur Research Network on Social Interactions and Economic Inequality. All the usual disclaimers apply. Correspondence: Department of Economics, University of California - Berkeley, 549 Evans Hall #3880, Berkeley, CA 94708. e-mail: *bgraham@econ.berkeley.edu.*

1 Introduction

Variation in many individual outcomes – such as earnings, academic achievement, substance abuse, criminal behavior, and technology adoption – includes a substantial between-group component. For example, a long economics of education literature documents that mean academic achievement varies dramatically across different classrooms, even among those located within the same school (e.g., Hanushek 1971). Perhaps the most straightforward explanation for this finding is the presence of classroom-level heterogeneity, such as differences in teacher quality.[1]

An alternative explanation for excess variance is that it mirrors the relative salience of social interactions or peer group effects. Social interactions are present if individual behavior is affected by reference or peer group behavior, characteristics or both. If students within the same classroom learn from one another, then achievement levels will covary positively within a classroom and hence display excess variation between classrooms. As with group-level heterogeneity, social interactions are associated with a lack of independence in outcomes across members of the same social group.

The two rival explanations for excess between-group variance, group-level heterogeneity and social interactions, are straightforward to understand, but exceptionally difficult to discriminate between empirically. Hoxby (2002, p. 58) emphasizes the "formidable obstacles" faced by researchers when attempting to detect peer effects in the learning process. In a recent and wide-ranging review Durlauf (2002, p. 20) concludes that "there is little reason why a skeptic should be persuaded to change his mind by the statistical evidence [on social interactions] currently available". Often associated with controversy, the empirical literature on social interactions is also characterized by widely divergent conclusions across different researchers.

The indecisiveness of available empirical evidence on social interactions partly reflects the fact that it speaks to some of the most contentious contemporary social and political issues in society. For example, the merits of school choice, ability tracking, busing and other desegregation measures, and different zoning laws all relate to the 'simple' question of whether peer group effects are important for the learning process.[2] A second reason for the diversity of conclusions found in the literature is that no consensus exists on how to best identify and estimate statistical models of social interactions in the presence of group-level heterogeneity.

This paper develops new methods for adducing the presence and magnitude of social interactions based on excess variance contrasts. An attractive feature of the proposed methods is that they are able to identify social interactions in a way that is robust to the presence of group-level heterogeneity. In the context of the economics of education example introduced above, they provide mechanisms by which excess between-classroom variation in student achievement can be decomposed into its teacher quality and peer effect portions. Such a decomposition is useful for assessing

[1] Another example is crime, which is endemic to some neighborhoods and negligible in other seemingly similar ones (c.f., Glaeser, Sacerdote and Scheinkman 1996). Piketty (2000) and Becker and Murphy (2000) survey theoretical models generating excess between-group variance.

[2] See Piketty (2000) for a related discussion.

the likely effects of various educational reforms, such as school choice or teacher accountability measures, on inequalities in student achievement.

The main ideas driving the formal identification results presented in this paper are easiest to explain within the context of the empirical application pursued below. The empirical application uses data from the Tennessee class size reduction experiment, Project STAR, to examine the effects of peers on early elementary school achievement levels. A key feature of Project STAR's experimental protocol is that it generated substantial variation in size across classrooms within the same school. Classrooms of two sizes are observed in the dataset: 'small' and 'large'. Students and teachers within participating schools were randomly assigned to one of the two types of classrooms.[3]

In this setting we would expect to observe more between-group variation in student ability across the set of small classrooms than across the set of large classrooms. In a large classroom any cluster of talented students will usually be offset by a corresponding cluster of below average students, resulting in a mean level of student ability that is similar across the set of large classrooms. In small classrooms, however, groups composed of mostly above or below average students are more likely to be observed, generating greater variation in mean ability. The differential variance in mean ability across the two types of classrooms induces a corresponding differential variance in attained achievement levels. Therefore a *mechanical* feature of the Project STAR data is greater between-group variation in academic achievement across small relative to large classrooms.

Now consider a difference in the between-group variances of achievement levels across small and large classrooms in the presence of unobserved variation in teacher quality. Since teachers were randomly assigned to either a small or large classroom, the distribution of teacher quality should be similar across the two types of classrooms. This implies that the difference in between-group variances across the two sets will be purged of the influence of any heterogeneity in teacher quality.

Peer effects generate positive outcome covariance, for example, because high ability students help other students learn more effectively.[4] Social interactions therefore amplify the mechanical difference in the between-group variance of academic achievement across small and large classrooms. Hence a ratio of the observed difference in between-group variances across small and large classrooms to an 'expected' difference provides a measure of the strength of social interactions, one free of the confounding influence of unobserved differences in teacher quality. To compute an estimate of the 'expected' difference in the between-group variance of academic achievement across small and large classrooms I use the within-group variation of the data, which is free from both the influence of social interactions and group-level heterogeneity.

The main idea is to exploit contrasts in excess variance across different types of classrooms to control for the effects of unobserved heterogeneity in teacher quality, a feature similar to panel data

[3] Section 2 provides a detailed overview of Project STAR, including a discussion of the consequences of likely deviations from the intended protocol for the identification strategies introduced below. The discussion here is heuristic only.

[4] Throughout I assume that peer effects are 'positive', with own outcomes increasing in peer outcomes and/or 'ability'.

analysis. The method also has a simple instrumental variables interpretation, which in addition to suggesting approaches to estimation and inference using standard software, also provides intuition about how identification works.

Section 2 reviews the relevant features of the Project STAR data which are used to illustrate the proposed methods. Section 3 provides a formal choice-theoretic motivation for the statistical model to which the identification results apply: the so-called linear-in-means model. The linear-in-means model was first analyzed by Manski (1993) and remains the workhorse of applied social interactions research, and for this reason it is a natural starting point for analyzing identification (c.f., Glaeser and Scheinkman 2003, Graham and Hahn 2004). The formal derivation of the statistical model highlights the key challenge of identifying social interactions: the problem of inferring the magnitude of individual responsiveness to changes in peer behavior when only equilibrium outcomes are observed. As in the econometrics of supply and demand, equilibrium plays a central role in understanding the nature of the observed data and in achieving identification. Empirical research that doesn't reflect a thorough understanding of the nature of choice and equilibrium in the presence of social interactions is unlikely to be persuasive.

Section 4 develops the main identification result of the paper and applies it to the Project STAR dataset. Throughout the Project STAR application illustrates how the results developed in the paper can be applied in a concrete and substantively interesting setting. This section also discusses what Manski (1993) dubbed the 'reflection problem'. The reflection problem has two distinct components.[5] The first component refers to the difficulty of distinguishing social interactions from group-level heterogeneity or *correlated* effects. This paper's main contribution is to show how conditional variance contrasts provide an innovative, intuitive, and attractive solution to this problem. The second component of the reflection problem is distinguishing *endogenous* social effects, where own behavior varies with mean peer group behavior, from *exogenous* social effects, where own behavior varies with predetermined peer characteristics. While this paper does not provide a solution to this problem, it does show how the reduced form estimate of the strength of social interactions derived below can be used to form relatively tight plausibility bounds on the magnitude of endogenous social effects.

Section 5 discusses the robustness of the identification strategy to various forms of misspecification as well as specific diagnostic tests. A method of moments interpretation of the main identification result indicates how standard omnibus specification tests can be used for diagnostic purposes. Unfortunately, omnibus tests typically lack power to detect certain directions of misspecification (Newey 1985). I therefore show how to assess robustness in two specific directions of misspecification that may be particularly salient in the Project STAR application: a lack of separability between teacher quality and class size in the educational production function and heterogenous class size effects. The analysis, while specific to the empirical application at hand,

[5] This typology follows directly from Proposition 1 and Corollary of Manski (1993, pp. 534 - 535).

illustrates how a combination of economic and statistical reasoning generates more powerful specification tests. Section 5 also discusses the implications of outcome variable measurement error. The results on measurement error are especially relevant for the student achievement application, since test scores provide only an imperfect measure of true attainment levels.

Section 6 exploits the experimental nature of Project STAR to compare standard best practice methods for estimating social interactions, based on excess between-group sensitivity, with estimation based on variance contrasts. In the excess sensitivity approach, exogenous variation in observed peer group composition, is used to implement simple regression-based tests for social interactions (e.g., Sacerdote 2001, Duncan *et al* 2003, Angrist and Lang 2004). A researcher applying these tests to the Project STAR dataset would be unable to reject the null of no social interactions. In contrast, this null is rejected using the excess variance methods developed in this paper. This apparent contradiction has a straightforward explanation related to test power.

Section 7 provides a formal characterization of the asymptotic power functions for the two types of tests. Calibrating these power functions to the Project STAR data confirms that excess variance tests are substantially more powerful in the current setting. Section 8 summarizes and outlines an agenda for further research.

The key contribution of this paper is to show how conditional variance contrasts can be used to discriminate excess between-group variance due to social interactions from that due to group-level heterogeneity. Relating social interactions to the presence of excess variance, however, is not a new idea, indeed it is typically excess variance or its cousin, positive residual covariance, which leads researchers to speculate that peer group effects may be present (e.g., Topa 1997, Gaviria 2000). Glaeser, Sacerdote and Scheinkman (1996), in a pioneering paper on crime patterns across U.S. cities, formally develop some of the connections between social interactions and excess between-group variance. In a follow-up paper, Glaeser and Scheinkman (2001) suggest 'scaling rules', based on city-size, that provide a partial solution to the problem of the confounding group-level heterogeneity and intuitively anticipate some of the results derived here. Solon, Page and Duncan (2000), building on ideas from the sibling and twins literature, also use analysis of covariance methods in an attempt to identify social interactions.

The key difference between the work of these authors and the approach advocated here is the use of conditional instead of unconditional covariances, resulting in full robustness to the presence of confounding group-level effects. Put differently, the results presented here identify the strength of social interactions as opposed to just bounding them. Casting the results into generalized method of moments form also provides a simple, convenient, and asymptotically valid framework for inference. More generally, using conditional variance restrictions to identify structural econometric models is uncommon in applied cross section econometrics research.[6]

[6] Rigobon (2004) is an interesting exception and also provides a historical review of the small literature in this area.

2 Brief Overview of Project STAR

Project STAR provides an ideal setting in which to test the identification strategy proposed in this paper since its experimental protocol generated exogenous variation in both class size *and* class composition. These two features of the data allow the excess variance approach developed below to be compared with standard best practice tests for social interactions based on excess between-group sensitivity.

Full details on Project STAR are provided by Finn *et al.* (2001), from which the following information is drawn. In the fall of 1985 entering kindergarten students in each of 79 project schools, located throughout the State of Tennessee, were randomly assigned to one of three class types within their school: small, with 15 to 17 students, regular, with 22 to 25 students, and regular with a full time teacher's aide, also with 22 to 25 students. I will often refer to these latter two types of classrooms as 'large'. Teachers were randomly assigned to classes in a second step. Schools participating in the project were required to be large enough to accommodate at least three kindergarten classes. Legislation also specified stratification across inner city, urban, suburban and rural schools.

During the first year 6,325 students, across 325 different classrooms, participated in the project. At the end of the year Stanford Achievement Tests in Mathematics and Reading were administered.[7] I have normalized the total scaled math and reading test scores by their sample mean and standard deviation. These normalizations make interpretation of the parameter estimates reported below straightforward.

Unfortunately the public release Project STAR dataset does not include a classroom identifier. However, using a simple algorithm based on grouping students with common values for school, class type (small, regular, or regular-with-aide) and teacher characteristics, I was able to uniquely assign 6,172 students to 317 classrooms; this sample is used in the remainder of the paper.[8]

Krueger (1999) provides a careful analysis of the STAR data, with an emphasis on assessing the salience of various threats to validity. His analysis indicates that the intended experimental protocol was carefully followed during the first year of the project. During subsequent years within-school variation in class type does not appear to be completely random due to a combination of deviations from protocol – in part due to parental pressure – and non-random attrition from the sample. For this reason the analysis presented in this paper is restricted exclusively to the kindergarten data.

The dataset includes four individual-level covariates: a dummy variable for student race (BLACK),

[7] No pre-intervention test scores are available.

[8] Boozer and Cacciola (2004) use a similar algorithm. Of the eight kindergarten classrooms excluded from the analysis two are regular classrooms and four are small classrooms which could not be individually separated; a further two classrooms were missing some teacher data and were also dropped. Twenty three kindergarten student records were missing information on free and reduced price school lunch eligibility, in these cases the missing values were replaced with either eligibility status for the same student in the closest of first, second or third grade (17 cases) or the median value among kindergarten students in their school (6 cases). In three cases missing student race values were replaced with school median values.

a dummy variable for gender (GIRL), a dummy variable for free or reduced price school lunch eligibility (FREELUNCH), and an age variable (DOB). The age variable is computed as the number of quarters after 1980 when a student was born. Positive values indicate a 'young' student, negative values 'older' students.

Classroom-level variables include: dummies for class type (SMALL, REGAIDE), number of students in a class (CLASSSIZE), dummies for teacher race and whether a teacher has a masters degree (BLACKTEACHER, MASTERS) and a years of teaching experience variable (EXPERIENCE). Also included is a career ladder variable indicating the current rank of a teacher on a six step scale (CLAD).[9] Since randomization only occurred within schools, all results reported below only use the within-school variation of the data.[10] Table 1 reports summary statistics for the individual test score and student characteristic data.

There is some ambiguity in descriptions of the experiment by original Project STAR researchers regarding whether students were randomly assigned to classrooms or only to class types (c.f., Finn *et al.* 2001). For 31 of the 79 participating schools this distinction is without content, since in those schools there were only 3 classrooms (one of each type). For the 48 schools with more than three classrooms students may have been non-randomly allocated across classrooms of the *same type*. For example, in larger schools administrators may have sought to balance the gender mix across such classrooms.

For the excess variance identification results presented below, the distinction between random assignment to classrooms and random assignment to class types is unimportant. For the traditional excess sensitivity estimator the distinction *is* important, with random assignment to classrooms required for identification. To assess whether the data are consistent with random assignment to classrooms I compute school-specific minimum χ^2 statistics for the null that the mean composition of a classroom, in terms of the four individual student characteristics discussed above, equals its corresponding school-wide mean.[11] Under the maintained null of random assignment to classrooms, the p-values associated with these statistics should be uniformly distributed.

If administrators mixed students according to an observed characteristic, we would expect the associated p-values to be left-skewed (i.e., classrooms look 'too alike'). Alternatively, if administrators stratified students, we would expect a right-skewed distribution, with classrooms looking 'too different'. Figure 1 plots histograms of the p-values for these tests with respect to BLACK, GIRL, FREELUNCH, and DOB respectively. There is little visual evidence of deviations from the uniform null distribution. More formally, Pearson χ^2 tests of the uniform null are easily accepted

[9] This variable is only available for 289 of the 317 teachers/classrooms in the core sample and hence is not used in most of the analysis.

[10] In practice this means that all the regression-based tests for social interactions include a full set of school dummy variables, while the excess variance tests work with residuals from a preliminary regression of test scores on the school dummies.

[11] See Vigdor and Nechyba (2004) for a discussion of this approach to testing for randomization.

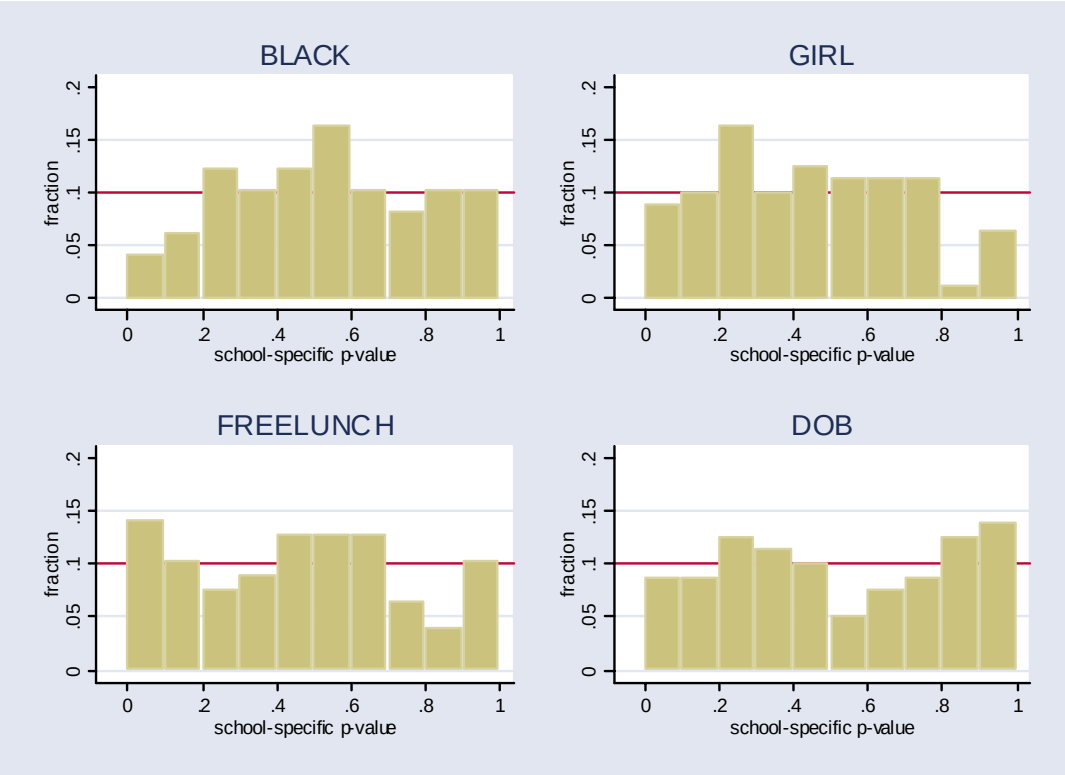

Figure 1: Plausibility Tests for Random Assignment to Classrooms
NOTES: The figure shows histograms of p-values associated with school-specific minimum χ^2 tests of the equality of composition across classrooms with respect to BLACK, GIRL, FREELUNCH, and DOB. The BLACK and FREELUNCH panels report p-values for the 49 and 78 schools with individual-level variation in these variables. The remaining two panels report p-values for all 79 project schools.

with p-values of 0.83, 0.24. 0.57 and 0.82 respectively.[12]

Not all kindergarten students report valid test score data. For the math test 5,724 students report valid scores and for the reading test 5,646 scores are valid (out of the 6,172 students in the core sample described above). Fortunately omissions of test scores appear to be idiosyncratic, in the sense that they are not predictable by any observable student, teacher or peer covariates. The analysis below assumes the pattern of missing test score data is indeed random and hence ignorable. Observed test scores are regarded as a random sample from different classrooms of known size. Importantly this case corresponds to the data structure most often available to economists interested in social interactions (e.g., a random subsample of individuals from different census tracts within a city as in Topa (2001)). Not observing all outcomes within a group does not complicate

[12] These tests are described by Cressie and Read (1984) among others. I discretize the p-value distribution by dividing the data into 10 equally-sized bins of width 0.1.

estimation. It does require modifying the relevant identifying moment conditions discussed below. Details of the required changes are provided in Appendix A.

3 The linear-in-means model of social interactions

This section sketches a simple model of choice in the presence of social interactions.[13] Individual utility depends on own attributes, a shared environment, and a group-specific stock of *social capital*. While individuals maximize utility treating social capital as fixed, its stock evolves endogenously with their choices. A *social equilibrium* is characterized by a mutually consistent vector of individual choices and stock of social capital. The endogeneity of social capital to the individual choices of group members makes it difficult to credibly identify social interactions.

We observe $c = 1, \ldots, N$ social, peer, or *reference groups* with the c^{th} group consisting of $i = 1, \ldots, M_c$ individuals. Conditional on group membership, individuals choose an action, y_{ci}, to maximize the indirect utility function

$$V\left(y_{ci}|\alpha_c, \varepsilon_{ci}, s_c\right) = -\frac{1-\xi}{2} y_{ci}^2 + \left(\alpha_c + \varepsilon_{ci}\right) y_{ci} - \frac{\xi}{2}\left(y_{ci} - s_c\right)^2, \tag{1}$$

where α_c represents group-level heterogeneity in institutions, prices, and other environmental factors shared by members of the same group, ε_{ci} represents individual-level heterogeneity in tastes arising from variation in income, family background, ability and so on, and s_c equals the group's stock of social capital. Utility depends on individual characteristics, shared environment, and social capital. Equation (1) is a version of the quadratic conformist utility function considered by Akerlof (1997) and, as will be seen below, is convenient for empirical work.[14] The first two terms of (1) capture what Akerlof terms intrinsic utility and the last term the disutility arising from individual deviations from community social norms, s_c, or extrinsic utility. High levels of social capital are complementary to the individual action, y_{ci}, which as emphasized by Becker and Murphy (2000), captures the idea that social forces may strongly influence individual behavior.

The preference structure given by (1) suggests that social interactions operate directly at the level of tastes. However, by viewing $(y_{ci}, \varepsilon_{ci}, \alpha_c, s_c)$ as inputs into a household production process, only the output of which households care about, any observed complementarity between own actions and social capital can be given a technological interpretation (Becker and Murphy 2000, p. 10). In this case (1) is a reduced form representation of household preferences. This interpretation is appropriate for the peer effects and student achievement application developed below.

Individuals treat the stock of social capital as fixed when choosing actions. Maximizing (1)

[13] Comprehensive reviews of the theory of choice in the presence of social interactions can be found in Becker and Murphy (2000), Brock and Durlauf (2001), and Glaeser and Scheinkman (2003).

[14] See also Brock and Durlauf (2001) and Glaeser and Scheinkman (2001, 2003).

yields an optimal action level of

$$y_{ci}^{BR}(\alpha_c, \varepsilon_{ci}, s_c) = \alpha_c + \xi s_c + \varepsilon_{ci}, \tag{2}$$

which is linear in social capital, s_c; $y_{ci}^{BR}(\alpha_c, \varepsilon_{ci}, s_c)$ is a function which maps $(\alpha_c, \varepsilon_{ci}, s_c)$ into individual choices. For any variable x_{ci} let $\overline{x}_c$ denote the group mean $M_c^{-1}\sum_{i=1}^{M_c} x_{ci}$ and $\underline{x}_c$ denote the vector $(x_{1c}, \ldots, x_{M_c})'$. Define the stock of social capital to be

$$s(\overline{y}_c, \overline{\varepsilon}_c) = \beta_* \overline{y}_c + \psi_* \overline{\varepsilon}_c.$$

Social capital is increasing in mean group actions, $\overline{y}_c$, and/or composition, $\overline{\varepsilon}_c$ (c.f., Manski 1993, Becker and Murphy 2000, Durlauf and Fafchamps 2004).

Substituting into (2) results in a modified best response function of

$$y_{ci}^{BR}(\alpha_c, \varepsilon_{ci}, s(\overline{y}_c, \overline{\varepsilon}_c)) = y_{ci}^{BR^*}(\alpha_c, \underline{\varepsilon}_c, \overline{y}_c) = \alpha_c + \beta \overline{y}_c + \psi \overline{\varepsilon}_c + \varepsilon_{ci}, \tag{3}$$

where $\beta = \xi\beta_*$ and $\psi = \xi\psi_*$; (3) determines agent $i's$ best response strategy for all *hypothetical* values of mean peer group behavior, $\overline{y}_c$, and composition, $\overline{\varepsilon}_c$. The goal is to estimate the parameters characterizing this reaction function $(\beta,\ \psi)$. Unfortunately the function is not observed, rather the data consist of only a single point on each individual reaction function. Analogous to supply and demand models, an equilibrium assumption will be a key component of addressing the implicit missing data problem and achieving identification (c.f., Manski 1995, Angrist, Graddy and Imbens 2000). Equation (3) defines the linear-in-means model of social interactions.[15]

3.1 Social equilibrium

A *social equilibrium* consists of a stock of social capital, s_c^e, and a $M_c \times 1$ vector of best responses strategies

$$\underline{y}_c^{BR}(\alpha_c, \underline{\varepsilon}_c, s_c^e) = \left(y_{c1}^{BR}(\alpha_c, \varepsilon_{ci}, s_c^e), \ldots, y_{cM_c}^{BR}(\alpha_c, \varepsilon_{ci}, s_c^e)\right)'$$

that are consistent with it

$$s_c^e = s\left(\sum\nolimits_{i=1}^{M_c} y_{ci}^{BR}(\alpha_c, \varepsilon_{ci}, s_c^e)/M_c, \overline{\varepsilon}_c\right). \tag{4}$$

In equilibrium all agents rationally anticipate the actions of their peers and choose best responses to those actions. In practice equilibrium is typically reached, not instantaneously, but rather through an adaptive learning process with individual updates in actions inducing changes in peer behavior in turn spurring further updates to own behavior. The resulting iterative 'reflection' process stops

[15] See Manski (1993), Brock and Durlauf (2001), Moffitt (2001). Graham and Hahn (2004) provide a methodologically-oriented review of the linear-in-means model.

when a (Nash) equilibrium is reached.

Let $y_{ci}^{e}(\alpha_c, \underline{\varepsilon}_c)$ denote the equilibrium action of the i^{th} individual:

$$y_{ci}^{e}(\alpha_c, \underline{\varepsilon}_c) = y_{ci}^{BR^*}(\alpha_c, \underline{\varepsilon}_c, \overline{y}_c^{e}(\alpha_c, \underline{\varepsilon}_c)).$$

Averaging over individuals within the same group and solving for $\overline{y}_c^{e}(\alpha_c, \underline{\varepsilon}_c)$ using (3) yields (assuming $\beta \neq 1$)

$$\overline{y}_c^{e}(\alpha_c, \underline{\varepsilon}_c) = \frac{\alpha_c}{1-\beta} + \frac{\psi+1}{1-\beta}\overline{\varepsilon}_c, \tag{5}$$

and hence, substituting (5) in (3), the reduced form

$$y_{ci}^{e}(\alpha_c, \underline{\varepsilon}_c) = \frac{\alpha_c}{1-\beta} + \frac{\psi+\beta}{1-\beta}\overline{\varepsilon}_c + \varepsilon_{ci}. \tag{6}$$

Henceforth, unless noted otherwise, I will assume that the observed vector of actions is an equilibrium vector, with y_{ci} denoting $y_{ci}^{e}(\alpha_c, \underline{\varepsilon}_c)$. In the absence of social interactions $y_{ci} = \alpha_c + \varepsilon_{ci}$, which is the standard one-way error component model.

3.2 A typology of social interactions

When $\beta \neq 0$ the marginal utility of individual action depends on peer actions; this dependence reflects the presence of *endogenous social interactions*. When β is greater (less) than zero there exist positive (negative) endogenous social interactions.[16] Positive interactions imply that the marginal utility of y_{ci} is increasing in peer action levels and hence that reaction functions slope upwards. When $\psi \neq 0$ each agent's chosen action level depends on the mean characteristics of her peers, reflecting the presence of *exogenous social interactions*. When $\alpha_c \neq 0$ correlated effects or *group-level heterogeneity* influence optimal action levels.[17, 18]

To better understand each of these three sources of variation in individual behavior it is helpful to consider the peer effects and student achievement example. Endogenous social effects capture the direct impact of peer achievement on own learning. These effects arise when students directly teach one another. They also can arise through agglomeration-type externalities. For example a teacher might be able to devote more attention to lagging students when most students have already mastered course material. In this case mean peer achievement affects individual achievement through its impact on available teacher time for own learning.

Exogenous or contextual effects arise when peer group background characteristics directly affect own learning. Survey evidence from the *Early Childhood Longitudinal Study* indicates that students from disadvantaged backgrounds begin kindergarten with lower levels of 'school readiness' (Lee and

[16] Cooper and John (1988) call these *strategic complementarity* and *substitutability* in agent actions respectively.

[17] This typology, which is now standard, was first formalized by Manski (1993).

[18] Brock and Durlauf (2001) and Glaeser and Scheinkman (2001, 2003) provide extended discussions of various types of interactive mechanisms.

Burkman 2002). To the extent that disadvantaged students are more disruptive or otherwise behave in ways not conducive to learning, peer socioeconomic background may directly alter one's learning environment and academic achievement.

Correlated or group effects arise because group members share a common environment. In an elementary school an obvious source of group-level heterogeneity is variation in teacher quality. Other sources might include various aspects of classroom quality (e.g., availability of natural light, background noise levels, comfort of chairs/desks etc.). In what follows I will, for expository reasons, equate group effects, α_c, with teacher quality, but it should be understood that this is only an approximation.

Determining the relative contributions of social interactions, group-level effects, and individual heterogeneity to variation in academic achievement has important policy implications. If social interactions are sizable then racial and socioeconomic residential segregation may be important for understanding inequality, and policies which focus on the distribution of peers across schools and classrooms may have first order effects on academic achievement. Alternatively, if group-level heterogeneity is relatively important, then resource distribution in the form of teacher quality and other school and classroom level inputs may be crucial. Finally, the relative contribution of individual-level heterogeneity, driven in turn by variations in family background and ability, may be envisioned as placing a bound on the efficacy of school-level policies to affect the distribution of academic achievement.

3.3 The social multiplier

The *social multiplier* provides a measure of the discrepancy between the initial response to a change in α_c or ε_{ci}, holding the stock of social capital fixed, and the full equilibrium response which occurs after all agents revise their strategies to new mutual best responses, a process which involves endogenous changes in the stock of social capital. In the presence of positive social interactions, full equilibrium responses typically exceed initial partial equilibrium responses, often substantially. The ratio of the full to partial equilibrium response equals the social multiplier, a useful summary measure of the strength of social interactions (c.f., Cooper and John 1988, Glaeser and Scheinkman 2003).

Formally we can define a social multiplier for changes in common environment, α_c, as well as for changes in group composition, $\underline{\varepsilon}_c$, as follows:

$$SM^{\alpha}\left(\alpha_c, \underline{\varepsilon}_c\right) = \frac{M_c^{-1}\sum_{i=1}^{M_c}\frac{\partial y_{ci}^{e}}{\partial \alpha_c}\left(\alpha_c, \underline{\varepsilon}_c\right)}{M_c^{-1}\sum_{i=1}^{M_c}\frac{\partial y_{ci}^{BR}(\alpha_c, \varepsilon_{ci}, s_c)}{\partial \alpha_c}|_{s_c = s(\overline{y}_c^{e}(\alpha_c, \underline{\varepsilon}_c), \overline{\varepsilon}_c)}} \tag{7}$$

$$SM^{\varepsilon}\left(\alpha_c, \underline{\varepsilon}_c\right) = \frac{M_c^{-1}\sum_{i=1}^{M_c}\frac{\partial y_{ci}^{e}}{\partial \varepsilon_{c1}}\left(\alpha_c, \underline{\varepsilon}_c\right)}{M_c^{-1}\sum_{i=1}^{M_c}\left(\frac{\partial y_{ci}^{BR}(\alpha_c, \varepsilon_{ci}, s_c)}{\partial \varepsilon_{ci}}|_{s_c = s(\overline{y}_c^{e}(\alpha_c, \underline{\varepsilon}_c), \overline{\varepsilon}_c)}\right)\cdot \mathbf{1}\left(i = 1\right)}. \tag{8}$$

The numerators of (7) and (8) measure the full equilibrium response, in terms of average group action levels, to changes in group environment, α_c, or a change in group composition, ε_{c1} (defining $SM^{\varepsilon}(\alpha_c, \underline{\varepsilon}_c)$ with respect to the first individual is done without loss of generality). The denominators are the mean partial equilibrium responses of agents to such changes.

From (5) and (6) it is straightforward to show that

$$SM^{\alpha}(\alpha_c, \underline{\varepsilon}_c) = \frac{1}{1-\beta}, \qquad SM^{\varepsilon}(\alpha_c, \underline{\varepsilon}_c) = \frac{\psi+1}{1-\beta}. \tag{9}$$

The size of the social multiplier for changes in *group environment* depends on reaction function slope: the more responsive individuals are to peer behavior the larger the multiplier. The size of the social multiplier for changes in *group composition* depends on reaction function slope as well as the strength of any exogenous or contextual effects. $SM^{\alpha}(\alpha_c, \underline{\varepsilon}_c)$ is a monotonic transformation of what Becker and Murphy (2000) call the social multiplier.[19] It is also identical to the social multiplier given by Cooper and John (1988) and Glaeser, Sacerdote and Scheinkman (2003). In the absence of exogenous effects, as is often assumed in applied work, $SM^{\alpha}(\alpha_c, \underline{\varepsilon}_c)$ and $SM^{\varepsilon}(\alpha_c, \underline{\varepsilon}_c)$ are identical.

To understand how the multiplier works we can decompose equilibrium changes in mean action levels due to changes in group composition into two components:

$$\begin{aligned} \frac{\partial \overline{y}^e_c(\alpha_c, \underline{\varepsilon}_c)}{\partial \varepsilon_{c1}} &= \frac{1}{M_c}\frac{\partial y^{BR}_{c1}(\alpha_c, \varepsilon_{c1}, s^e_c)}{\partial \varepsilon_{c1}} \\ &\quad + \frac{1}{M_c}\sum\nolimits_{i=1}^{M_c} \frac{\partial y^{BR}_{ci}(\alpha_c, \varepsilon_{ci}, s^e_c)}{\partial s^e_c}\left[\frac{\partial s^e_c}{\partial \overline{y}^e_c}\frac{\partial \overline{y}^e_c(\alpha_c, \underline{\varepsilon}_c)}{\partial \varepsilon_{c1}} + \frac{\partial s^e_c}{\partial \overline{\varepsilon}_c}\frac{\partial \overline{\varepsilon}_c}{\partial \varepsilon_{c1}}\right]. \end{aligned} \tag{10}$$

The first component captures the direct effect of changes in ε_{c1} on mean group behavior *holding social capital constant.* This is the private effect of changes in own characteristics on own behavior. Individual shocks also affect group behavior through their impact on the equilibrium stock of social capital. They alter social capital through the endogenous and exogenous effects. Endogenous social interactions generate a feedback effect whereby initial changes in agent behavior induce further changes in peer behavior and *vice versa* until a new equilibrium is reached. To see this solve (10) for $\partial \overline{y}^e_c(\alpha_c, \underline{\varepsilon}_c)/\partial \varepsilon_{c1}$ to get

$$\frac{\partial \overline{y}^e_c(\alpha_c, \underline{\varepsilon}_c)}{\partial \varepsilon_{c1}} = \frac{1}{M_c}\frac{\frac{\partial y^{BR}_{c1}(\alpha_c, \varepsilon_{c1}, s^e_c)}{\partial \varepsilon_{c1}} + \sum_{i=1}^{M_c}\frac{\partial y^{BR}_{ci}(\alpha_c, \varepsilon_{ci}, s^e_c)}{\partial s^e_c}\frac{\partial s^e_c}{\partial \overline{\varepsilon}_c}\frac{\partial \overline{\varepsilon}_c}{\partial \varepsilon_{c1}}}{1 - M_c^{-1}\sum_{i=1}^{M_c}\frac{\partial y^{BR}_{ci}(\alpha_c, \varepsilon_{ci}, s^e_c)}{\partial s^e_c}\frac{\partial s^e_c}{\partial \overline{y}^e_c}}.$$

Endogenous social interactions amplify the effects of shocks to group composition on outcomes (c.f., Glaeser, Sacerdote and Scheinkman 1996, Becker and Murphy 2000). This observation is

[19] In particular, Becker and Murphy (2000) refer to β – the average slope of individual agent reaction functions – as the social multiplier.

central to the excess variance intuition which drives the identification results given below.

4 Identification and estimation

The data consist of a sample n individual *equilibrium* outcomes across N groups ($n = \sum_{c=1}^{N} M_c$). All individuals within a group are sampled. Extending what follows to the case where only a random subset of outcomes are observed within each group is straightforward but, although important for empirical applications, needlessly complicates the development of the main results.[20] Group size, M_c, is distributed multinomially with support $\mathcal{M} \equiv \{m_1, \ldots, m_S\}$. We also observe a vector of instruments, q_c, which vary across groups; their precise role will be made clear below.

A naive approach to detecting social interactions attempts to estimate β by a least squares regression of y_{ci} on $\overline{y}_c$. This is a tautological regression and, with a little introspection, it is obvious that $\widehat{b} = 1$, which will generally differ from β.

The reduced form variance-covariance matrix of equilibrium outcomes, however, is identified. Let $w_c = (M_c, q_c')'$. Assume that the conditional mean and variance of $(\underline{\varepsilon}_c', \alpha_c)'$ equal

$$E\left[\underline{\varepsilon}_c', \alpha_c | w_c\right] = \left(0 \cdot \iota_{M_c}', \ \mu\left(w_c\right)\right), \tag{11}$$

with ι_{M_c} denoting an $M_c \times 1$ vector of ones, and

$$Var\left(\underline{\varepsilon}_c', \alpha_c | w_c\right) = \begin{pmatrix} \sigma^2\left(w_c\right) & \sigma_{\varepsilon\varepsilon}\left(w_c\right) & \cdots & \sigma_{\varepsilon\varepsilon}\left(w_c\right) & \sigma_{\alpha\varepsilon}\left(w_c\right) \\ \sigma_{\varepsilon\varepsilon}\left(w_c\right) & \sigma^2\left(w_c\right) & & \vdots & \vdots \\ \vdots & & \ddots & \sigma_{\varepsilon\varepsilon}\left(w_c\right) & \vdots \\ \sigma_{\varepsilon\varepsilon}\left(w_c\right) & \cdots & \sigma_{\varepsilon\varepsilon}\left(w_c\right) & \sigma^2\left(w_c\right) & \sigma_{\alpha\varepsilon}\left(w_c\right) \\ \sigma_{\alpha\varepsilon}\left(w_c\right) & \cdots & \cdots & \sigma_{\alpha\varepsilon}\left(w_c\right) & \sigma_{\alpha}^2\left(w_c\right) \end{pmatrix}, \tag{12}$$

respectively.

Condition (11) is unrestrictive. Condition (12) is also weak but merits some discussion. It implies that, conditional on w_c, equilibrium outcomes across members of the same social group are equicorrelated. This seems well-motivated by standard (finite) exchangeability considerations. Looking within a given classroom there is no *a priori* reason to assume, for example, that the 1^{st} student is 'brighter' than the 10^{th} or that the 2^{nd} student is more like the 3^{rd} than the 4^{th} student (c.f, Rubin 1981). Similarly, conditional on class size and the instrument, there is no *a priori* reason to think achievement in the 23^{rd} classroom should be higher or lower than in 53^{rd} (assuming both classrooms have identical values of w_c).

[20] Appendix A details how to implement the random subsample case.

Averaging across group size but continuing to condition on q_c we define the notation

$$E\left[\sigma^2(w_c)|q_c\right] = \sigma^2(q_c), \quad E\left[\sigma_{\varepsilon\varepsilon}(w_c)|q_c\right] = \sigma_{\varepsilon\varepsilon}(q_c),$$
$$E\left[\sigma_{\alpha\varepsilon}(w_c)|q_c\right] = \sigma_{\alpha\varepsilon}(q_c), \quad E\left[\sigma_\alpha^2(w_c)|q_c\right] = \sigma_\alpha^2(q_c), \quad E\left[\mu(w_c)|q_c\right] = \mu(q_c).$$

Averaging over both M_c and q_c we have

$$E\left[\sigma^2(w_c)\right] = \sigma^2, \quad E\left[\sigma_{\varepsilon\varepsilon}(w_c)\right] = \sigma_{\varepsilon\varepsilon}, \quad E\left[\sigma_{\alpha\varepsilon}(w_c)\right] = \sigma_{\alpha\varepsilon}, \quad E\left[\sigma_\alpha^2(w_c)\right] = \sigma_\alpha^2, \quad E\left[\mu(w_c)\right] = \mu.$$

From (11) and (12), we have $\Omega(w_c) \equiv Var(\underline{y}_c|w_c)$ equal to

$$\begin{aligned} &\sigma^2(w_c)(1-\zeta_{\varepsilon\varepsilon}(w_c))I_{M_c} + \sigma^2(w_c)\zeta_{\varepsilon\varepsilon}(w_c)\iota_{M_c}\iota'_{M_c} + \\ &\left\{\frac{\sigma_\alpha^2(w_c) + 2(\psi+1)\sigma_{\alpha\varepsilon}(w_c)}{(1-\beta)^2} + \left(\gamma^2-1\right)\sigma^2(w_c)\left[\zeta_{\varepsilon\varepsilon}(w_c) + \frac{1-\zeta_{\varepsilon\varepsilon}(w_c)}{M_c}\right]\right\}\iota_{M_c}\iota'_{M_c} \end{aligned} \tag{13}$$

where I_{M_c} is an $M_c \times M_c$ identity matrix, $\zeta_{\varepsilon\varepsilon}(w_c) = \sigma_{\varepsilon\varepsilon}(w_c)/\sigma^2(w_c)$ denotes the conditional correlation of individual-level characteristics, ε_{ci}, across members of the same group, and $\gamma = (\psi+1)/(1-\beta)$ is the social multiplier defined in equation (9) above.

Equation (13) illustrates how the conditional variance of equilibrium outcomes consists of four distinct components:

1. INDIVIDUAL HETEROGENEITY $\sigma^2(w_c)$: Outcome variance depends on heterogeneity in individual-level background characteristics, ε_{ci}.

2. ENVIRONMENTAL/GROUP HETEROGENEITY $\sigma_\alpha^2(w_c)$: Outcome variance depends on heterogeneity in shared environment across social groups, α_c.

3. MATCHING $\sigma_{\alpha\varepsilon}(w_c)$: Individuals' background characteristics may be correlated with the group's common environment (i.e., $cov(\alpha_c, \underline{\varepsilon}_c) \neq 0$). This affect arises if individuals select on α_c when choosing a group or if α_c varies endogenously in response to changes in group composition.[21] Positive matching – $\sigma_{\alpha\varepsilon}(w_c) > 0$ – amplifies the between-group component of outcome variance.

4. GROUP COMPOSITION $\sigma^2(w_c)\left[\zeta_{\varepsilon\varepsilon}(w_c) + (1-\zeta_{\varepsilon\varepsilon}(w_c))M_c^{-1}\right]$: The variance of individual outcomes depends on heterogeneity in mean group-composition, $\overline{\varepsilon}_c$. Variance in $\overline{\varepsilon}_c$ across groups depends on the degree of sorting, group-size and the amount of individual-level heterogeneity in the population (i.e., $\zeta_{\varepsilon\varepsilon}(w_c)$, M_c, and $\sigma^2(w_c)$ respectively). Increased sorting implies greater correlation of individual attributes across members of the same group and

[21] For example, a teacher's effort might change with the average ability of her students.

hence, for any given population-wide distribution of individual heterogeneity, more between-group variation in $\overline{\varepsilon}_c$. In the presence of social interactions ($\gamma \neq 0$) the *variance* of individual outcomes depends on the *variance* of group-composition.

Since $\Omega(w_c)$ has a one-way error component structure it is sufficient to work directly with within- and between- transforms of $\underline{y}_c$ as opposed to the entire $M_c \times M_c$ matrix. Consider the following two within- and between-group transforms of the data:

$$g_c^b = (\overline{y}_c - \mu(q_c))^2, \qquad g_c^w = M_c^{-1}(M_c - 1)^{-1} \sum_{i=1}^{M_c} (y_{ci} - \overline{y}_c)^2. \tag{14}$$

Taking expectations conditional on the instrument, q_c, we have

$$E\left[g_c^b | q_c\right] = \varsigma(q_c) + \gamma^2 E\left[\frac{\sigma^2(w_c)(1 - \zeta_{\varepsilon\varepsilon}(w_c))}{M_c} | q_c\right], \tag{15}$$

$$E\left[g_c^w | q_c\right] = E\left[\frac{\sigma^2(w_c)(1 - \zeta_{\varepsilon\varepsilon}(w_c))}{M_c} | q_c\right], \tag{16}$$

where

$$\varsigma(q_c) \equiv \frac{\sigma_\alpha^2(q_c) + 2(\psi + 1)\sigma_{\alpha\varepsilon}(q_c) + (\psi + 1)^2 \sigma_{\varepsilon\varepsilon}(q_c)}{(1 - \beta)^2}.$$

Equation (15) and (16) illustrate how group-level heterogeneity, matching, sorting, and social interactions manifest themselves differently in the within- and between-group variation of the data.

A key insight of this paper is that the within-group variation of the data provides information on the amount of between-group variance we would expect to observe in the absence of social interactions. I will refer to $E[g_c^w | q_c]$ as 'expected' between-group variance. By 'expected' I mean the portion of between-group variance that is estimable from the within-group variation of the data alone. Loosely speaking, $E[g_c^w | q_c]$ equals the between-group variance in outcomes we would expect to observe in the absence of group-level heterogeneity, matching, sorting, and social interactions.

Under the two the identifying assumptions

$$\sigma_\alpha^2(q_c) \equiv \sigma_\alpha^2, \qquad \sigma_{\alpha\varepsilon}(q_c) \equiv \sigma_{\alpha\varepsilon}, \qquad \sigma_{\varepsilon\varepsilon}(q_c) \equiv \sigma_{\varepsilon\varepsilon} \tag{17}$$

$$E\left[\sigma^2(M_c, q_c) M_c^{-1} | q_c = q\right] \neq E\left[\sigma^2(M_c, q_c) M_c^{-1} | q_c = q'\right], \qquad q = q' \tag{18}$$

the conditional moment

$$E[\rho(g_c, \theta) | q_c] = 0 \tag{19}$$

identifies $\theta = \left(\varsigma, \gamma^2\right)'$ with $\rho(g_c, \theta) = g_c^b - \varsigma - \gamma^2 g_c^w$ and $g_c = \left(g_c^w, g_c^b\right)'$.[22]

[22] Under assumption (17) we can, without loss of generality, redefine ε_{ci} and α_c to be orthogonal by setting

To motivate (17) and (18) it is instructive to return to the Project STAR application. Let q_c equal a class type indicator, taking a value of 1 if the c^{th} classroom is of the small type and zero otherwise. Condition (17) states that the variance of group-level heterogeneity as well as the intensity of matching and sorting are equal across small and large classrooms. Both students and teachers were randomly assigned to either small or large kindergarten classrooms as part of Project STAR and for this reason (17) seems reasonable to impose.

Random assignment to classrooms ensures the absence of any matching or sorting, implying that $\sigma_{\alpha\varepsilon}(q_c) = \sigma_{\varepsilon\varepsilon}(q_c) \equiv 0$. If random assignment is to class types only, then identification requires that any within-class-type matching and/or sorting does not operate differentially across small and large classrooms.[23] While random assignment of teachers does not eliminate heterogeneity in teacher quality, it does imply that the distribution of observed and unobserved teacher characteristics across classrooms should be independent of class type. Therefore, as long as class type and latent teacher quality are separable in the educational production function, random assignment implies that $\sigma^2_\alpha(q_c)$ is constant in q_c. This assumption is discussed further in Section 5 below.

Condition (18) states that the expected between-group variance across the two sets of classrooms differs. Variation in mean group-size across the two sets will typically be sufficient to ensure that (18) holds and, in any case, the assumption is straightforward to test.

With q_c binary (19) is equivalent to the unconditional moment $E\left[(1, q_c)' \rho(g_c, \theta)\right] = 0$ which, directly solving for γ^2, yields the Wald-IV estimate

$$\gamma^2 = \frac{E\left[g_c^b | q_c = 1\right] - E\left[g_c^b | q_c = 0\right]}{E\left[g_c^w | q_c = 1\right] - E\left[g_c^w | q_c = 0\right]}. \tag{20}$$

The structure of (20) provides insight into how identification works. The numerator is a contrast of observed or actual between-group variances across small and large classrooms. Under (17) this contrast will not be influenced by the variance of group-level heterogeneity, matching, and sorting, since these three effects are constant across the two sets of groups and hence are differenced away in the numerator. The numerator will only reflect *differences* in the variance of mean individual-level heterogeneity, $Var(\overline{\varepsilon}_c | q_c)$, across the two sets of groups, *as amplified by social interactions.*

The denominator also equals the difference in the variance of mean individual-level heterogeneity, but unamplified by social interactions. The sample analog of the ratio of the two differences therefore provides a consistent estimate of γ^2, the square of the social multiplier with respect to changes in group composition. Equation (20) also makes clear that assumptions (17) and (18) are simply special versions of the standard exclusion and rank restrictions required for a valid

$\varepsilon^*_{ci} = \varepsilon_{ci} - E^*[\varepsilon_{ci} | \alpha_c]$ and $\alpha^*_c = \alpha_c + E^*[\varepsilon_{ci} | \alpha_c]$ respectively, where $E^*[\cdot]$ denotes a linear predictor as in Chamberlain (1984). This transformed group-effect, α^*_c, reflects a combination of 'true' group-level heterogeneity and matching. I nonetheless continue to work with the orginal parameterization to emphasize the independent contributions of these forces.

[23] As an example of differential sorting consider the case where administrators seek to mix according to unobserved ability across small classrooms while stratifying along the same dimension in large classrooms.

instrumental variable in the linear simultaneous equations model.

4.1 Excess variance contrasts in Project STAR

Project STAR generated substantial exogenous variation in class size. As discussed in the introduction, under random assignment of students to class type, $\overline{\varepsilon}_c$ – say mean 'ability' – will vary more across the set of small classrooms than across the set of large ones. Formally, if small classrooms are all of size of m_S and large ones all of size $m_L > m_S$, then the ratio of the variance of mean ability in small to large classrooms will be $m_L/m_S > 1$.[24] For relatively small values of m_S and m_L this ratio can be substantial. In the Project STAR dataset the mean size of small kindergarten classrooms is about 15, while the mean size of regular and regular-with-aide classrooms is about 22, suggesting that the variance of mean ability should be approximately 1.5 times greater across small classrooms than large. This contrast provides a powerful mechanism with which to identify social interactions.

Table 2 reports evidence of large differences in the variance of mean test scores and class composition by class type. For example, the variance in mean classroom math and reading test scores is 1.8 and 1.9 times greater in small versus large classrooms (column 4). Similarly heterogeneity in mean socioeconomic status, gender mix, and age structure are significantly greater in small classrooms.[25] Clearly small classrooms are more heterogeneous in terms of mean composition than large classrooms, consistent with the requirements of (18).

In contrast to mean ability, under (17) the dispersion of teacher quality, intensity of student-teacher matching, and amount of sorting should not vary with class type. These assumptions may at first glance appear somewhat abstract, but they are relatively straightforward to assess. Verifying their plausibility requires consideration of the mechanism by which students and teachers are assigned to class types. In the case of Project STAR random assignment of teachers to class type ensures that variation in teacher quality should be similar across small and large classrooms within the same school. Random assignment of students to class type suggests that any correlation of ability across students within the same classroom or between student ability and teacher quality should also be constant in class type.[26]

Tables 2, 3, and 4 assesses these claims using the limited number of student and teacher characteristics available in the Project STAR public release data. In contrast to the individual-level characteristics, the bottom portion of Table 2 shows that variation in observed teacher characteristics is not significantly different across small and large classrooms for three of the four teacher

[24] This calculation makes the simplifying assumption that student ability, ε_{ci}, is homoscedastic with respect to class type.

[25] That the variance of racial composition is not significantly different across the two types of classrooms reflects the lack of within-school variation in race in the Project STAR data (30 of the 79 schools have either all black or all white kindergartens; the majority of the balance have only a handful of either black or white students).

[26] This assumes that any within-class-type sorting and/or matching of students with teachers is equally intense in small and large classrooms.

variables.

Table 3 examines the pattern of covariance between student and teacher characteristics in small and large classrooms. The null hypothesis of an equal covariance for each of the 16 pairs of student and teacher characteristics across class types is accepted in all cases. There is no evidence of a differential pattern of student-teacher matching in small versus large classrooms, at least based on the characteristics available in the Project STAR data. Indeed there is no evidence of student-teaching matching at all.

Table 4 calculates the conditional covariance of observed characteristics across students within the same classroom. There is modest evidence of mixing or balancing on the part of school administrators. For example, race is slightly negatively correlated across students within the same classroom. This effect is small and, most importantly, there is no evidence of *differential* mixing across small and large classrooms.

While the primary motivation for assuming (17) and (18) is the experimental structure of Project STAR, tables 2, 3, and 4 provide substantial auxiliary evidence of their plausibility.

Table 5 reports estimates of γ^2 using the Project STAR kindergarten Stanford Achievement Test scores in mathematics and reading and the sample analog of (20). The sample used consists of 6,172 students from 317 classrooms, 123 of which are small with the remaining 194 being large. Implementation requires forming g_c^b and g_c^w using the formulae given in (14). In the case of g_c^b this requires replacing $\mu(q_c)$ with a consistent estimate. Since the support of q_c is binary $\mu(q_c)$ is straightforward to estimate by a least squares regression. For continuous valued instruments $\mu(q_c)$ can be replaced with any of a number of nonparametric estimates.

In addition to removing conditional mean heterogeneity across class types I orthogonalize test scores with respect to a matrix of school dummy variables; this reduces the amount of group-level heterogeneity in the data, improving precision. Specifically residuals, $\widehat{u}_{ci}$, from a preliminary regression of normalized test scores on a matrix of school dummies and the class type instrument, q_c, are used to compute $g_c^b = \overline{\widehat{u}}_c^2$ and $g_c^w = M_c^{-1}(M_c - 1)^{-1} \sum_{i=1}^{M_c} (\widehat{u}_{ci} - \overline{\widehat{u}}_c)^2$. It is straightforward to show that the asymptotic distribution of the Wald estimator is unaffected by this preliminary step. No correction to the conventional heteroscedastic robust standard errors reported by a basic regression program needs to be made (c.f., MaCurdy 1982).[27]

The first row of Panel A of Table 5 computes the mean between-group variance in math test scores for small and large classrooms. As shown in column 3 there are significant differences in the amount of between-group variance across the two sets of classrooms. The difference in the two variance terms equals the numerator of the Wald estimator. Row 2 of the table computes 'expected' between-group variance measures for small and large classrooms using the within-group variation of the data. Here too there are significant differences across the two types of classrooms. The difference in expected between-group variances is statistically significant (row 2, column 3). The

[27] The higher order properties of $\widehat{\gamma}^2_{WALD}$ will be affected as discussed by Newey, Ramalho and Smith (2005).

F-Statistic associated with this difference equals the standard 'first-stage' diagnostic advocated by Staiger and Stock (1997); this statistic is reported in row 4. The first stage F-statistics equal 51.01 and 16.27 for math and reading respectively. By the standards of the classical linear simultaneous equations model with homoscedastic normal errors, these F-statistics suggest strong identification (c.f., Stock, Wright and Yogo 2002, Stock and Yogo 2004). Although these results do not apply directly here, since the structural and first-stage errors are both heteroscedastic and right skewed, I tentatively conclude that identification is strong enough to obviate weak instrument concerns, an assessment that is subjected to more rigorous scrutiny in Graham (2004).

The Wald estimate, $\widehat{\gamma}^2_{WALD}$, is the ratio of the row 1 and 2 column 3 differences; this ratio is reported in row 3.[28] The Wald estimate for γ^2 indicates that the *difference* in between-group variance across small and large classrooms is over three times what we would expect in the absence of social interactions. Panel B repeats the exercise for reading achievement test scores. In this case excess variance contrasts are almost four times what we would expect in the absence of social interactions. These point estimates suggest social multipliers of 1.76 and 1.97 for math and reading achievement respectively.

Table 6 reports three tests of the no social interactions null (i.e., $\gamma = 1$). The first test, reported Panel A, is a conventional Wald test for the null that $\gamma^2 = 1$. It is based on the standard normal approximation to the sampling distribution of $\widehat{\gamma}^2_{WALD}$ and is easy to compute; unfortunately it has poor power properties in the current setting. Observe that g^b_c and g^w_c are functions of squared outcomes, generating a right-skewed sampling distribution for the identifying moment function, $\rho(g_c, \theta)$. In finite samples this right-skewness also manifests itself in the sampling distribution of $\widehat{\gamma}^2_{WALD}$. Since the Wald test is based on a symmetric asymptotic approximation to the sampling distribution of $\widehat{\gamma}^2_{WALD}$, it behaves poorly when the actual (finite) sampling distribution is asymmetric. In particular Wald confidence intervals, symmetric by design, will extend too far to the left and too little to the right of $\widehat{\gamma}^2_{WALD}$, suggesting low power for a Wald test of the no social interactions null, in addition to possible size distortion.

One approach to dealing with this problem is to test the null $\gamma = 1$ directly, again using a Wald test, with the appropriate variance calculated via the delta method. Intuitively, we would expect the *square root of* $\widehat{\gamma}^2_{WALD}$ to have a more symmetric sampling distribution and, consequently, greater accuracy of the large sample normal approximation. An additional advantage of this approach is that γ is the actual parameter of interest. Panel B of the table reports p-values for the no social interactions null based on this test as well as corresponding confidence intervals for $\widetilde{\gamma} = \sqrt{\widehat{\gamma}^2_{WALD}}$. This test results in a more decisive rejection of the null hypothesis of no social interactions.

The difference between Panels A and B stems from the Wald statistic's lack of invariance to one-to-one transformations of the null hypothesis. While intuition may privilege one statement of

[28] This ratio is easily computed using standard software by calculating the instrumental variables regression of g^b_c on 1 and g^w_c where g^w_c is instrumented using q_c (the class type dummy variable). The heteroscedastic robust standard errors reported in the regression output are also asymptotically valid.

the null over another, this lack of invariance is unattractive. The third test, reported in Panel C of the table, addresses this shortcoming. It is based on the asymptotic approximation to the sampling distribution of the profiled Empirical Likelihood (EL) saddle-point criterion function. This statistic is invariant to parameter transformations and is, essentially, a method of moments analog to a log-likelihood ratio test (c.f., Owen 2001, Newey and Smith 2004). Confidence intervals based on this test need not be symmetric and can consequently better capture asymmetric uncertainty surrounding $\widehat{\gamma}^2_{WALD}$ when its sampling distribution is skewed. Graham (2005) provides a more complete discussion of this test as well as supporting Monte Carlo evidence calibrated to mimic the Project STAR dataset. The p-values for the no social interactions null based on this test are in between those given by the two Wald tests. Confidence intervals for both $\widehat{\gamma}^2_{WALD}$ and $\widetilde{\gamma} = \sqrt{\widehat{\gamma}^2_{WALD}}$ based on inverting the test are also reported in panel C. The EL intervals are both longer than the panel A and B Wald intervals and right skewed, indicating more uncertainty above the point estimate than below; this is particularly the case for the reading estimates, which are less strongly identified.

To summarize, the estimates of γ^2 reported in Table 5 suggests a social multiplier of between 1.07 and 2.31 for math achievement with a point estimate of 1.76, and one between 1.05 and 3.07 with a point estimate of 1.97 for reading achievement. These ranges are based on the 95 percent EL confidence intervals reported in Table 6.

These confidence intervals generally encompass other recent estimates based on different samples, grade levels, and methods. For example, work by Angrist and Lang (2004), using elementary school test score data from the Metco program in Brookline, MA, implies a point estimate for the social multiplier of about 1.25 (albeit insignificantly different from 1). Lefgren (2004), using elementary school data from the Chicago Public schools system, reports a statistically significant best estimate of β consistent with a very small social multiplier of about 1.05. Hoxby's (2002) results, using data from the Texas Schools Project, are consistent with values in the middle to the high end of the ranges given above. Boozer and Cacciola (2004), also using Project STAR data but a different identification strategy, report estimates of β consistent with social multipliers in excess of 10, well outside the 95 percent confidence intervals for the estimates given here.

4.2 The reflection problem: endogenous or exogenous social interactions?

Manski's (1993) reflection problem consists of two parts. The first part is discriminating social interactions from group-level heterogeneity. Contrasts in excess variance provide one solution to this problem. The second part of the reflection problem is distinguishing *endogenous* social effects ($\beta \neq 0$), where own behavior varies with mean peer group behavior, from *exogenous* social effects ($\psi \neq 0$), where own behavior varies with predetermined peer characteristics. While contrasts in excess variance are unable to distinguish between the two types of social effects, the estimated value of γ^2 can easily be combined with subjective prior information to construct Bayesian credibility

sets for β.

The endogenous effect parameter, β, can be written as a function of the social multiplier, γ, and the exogenous effects parameter, ψ,

$$\beta = 1 - \frac{\psi + 1}{\sqrt{\gamma^2}}.$$

From a joint posterior distribution for $\left(\gamma^2, \psi\right)$ we can therefore calculate a posterior distribution for β. This posterior can be used to assess the evidence for or against endogenous social interactions.

Appealing to the Bernstein-von Mises Theorem yields an approximate posterior distribution for γ^2 of

$$\overline{\pi}\left(\gamma^2|z\right) \overset{D}{\simeq} \mathcal{N}(\widehat{\gamma}^2, se^2_{\widehat{\gamma}^2}),$$

where $se_{\widehat{\gamma}^2}$ is the estimated large sample standard error of the Wald estimate, $\widehat{\gamma}^2$, and z denotes all available data. I assume that sample information dominates any available prior information about γ^2.

Unfortunately, data do not provide any information with which to update priors regarding the magnitude of ψ. This reflects the lack of separate identification of β and ψ via excess variance contrasts. To deal with non-identification I place an informative prior on ψ which, in the absence of sample information, also equals its posterior (i.e., $\overline{\pi}\left(\psi|z\right) = \pi\left(\psi\right)$).[29]

Assuming independence of the priors (and hence the posteriors) for γ^2 and ψ, an application of the change-of-variables formula yields a joint posterior pdf for (β, ψ) of

$$\overline{\pi}\left(\beta, \psi|z\right) = \frac{2\left[\gamma^2\left(\beta, \psi\right)\right]^{(3/2)}}{\psi + 1}\overline{\pi}\left(\gamma^2\left(\beta, \psi\right)|z\right)\overline{\pi}\left(\psi|z\right),$$

recalling that $\gamma^2\left(\beta, \psi\right) = \left(\frac{\psi+1}{1-\beta}\right)^2$.[30] Integrating out ψ yields the marginal posterior density of β,

$$\overline{\pi}\left(\beta|z\right) = \int_{\psi\in\Psi} \frac{2\left[\gamma^2\left(\beta, \psi\right)\right]^{(3/2)}}{\psi + 1}\overline{\pi}\left(\gamma^2\left(\beta, \psi\right)|z\right)\overline{\pi}\left(\psi|z\right)d\psi. \tag{21}$$

[29] Most researchers achieve identification by using the exceptionally informative 'dogmatic prior' of $\psi = 0$.

[30] The determinant of the Jacobian needed for the change-of-variables formula is

$$\left|\begin{array}{cc} \frac{1}{2}\frac{\psi+1}{\left[\gamma^2\right]^{(3/2)}} & -\sqrt{1/\gamma^2} \\ 0 & 1 \end{array}\right| = \frac{1}{2}\frac{(\psi+1)}{\left[\gamma^2\right]^{(3/2)}}.$$

Note that $\frac{2\left[\gamma^2(\beta,\psi)\right]^{(3/2)}}{\psi+1}\overline{\pi}\left(\gamma^2\left(\beta,\psi\right)|z\right)$ is identical to the pdf of a normal random variable with mean $\frac{1}{2}\frac{(1+\psi)}{\left[\widehat{\gamma}^2\right]^{(3/2)}}$ and variance $\frac{1}{4}\frac{(1+\psi)^2}{\left[\widehat{\gamma}^2\right]^3}\cdot se^2_{\widehat{\gamma}^2}$; a delta approximation to the sampling distribution of $\widehat{\beta} = 1 - \frac{\psi+1}{\sqrt{\widehat{\gamma}^2}}$ (with ψ known) yields an identical distribution.

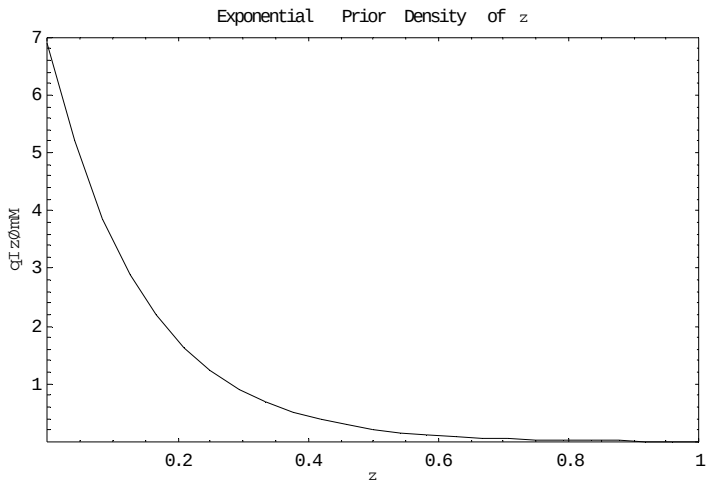

Figure 2: Prior Density of ψ

Implementing this approach requires placing an informative prior on ψ. This can be done by considering a range of plausible values for ψ. In the present setting introspection suggests that assuming $\psi \in [0, 1]$ is reasonable; this implies that able students make for good peers ($\psi \geq 0$) but that the direct effect of own ability on own achievement is at least as large as the direct effect of mean peer ability on own achievement ($\psi \leq 1$). How to distribute the probability density on 0 to 1 interval is less clear. Here I assume that ψ is an exponential random variable, where the scale parameter, λ, is chosen such $\int_{\psi=1}^{\infty} \pi(\psi; \lambda)\, d\psi = 0.001$; this ensures that 99.9 percent of the prior probability is placed on the 0 to 1 interval.[31] This prior embodies the subjective notation that any exogenous effects are likely to be small relative to the importance of own ability; it implies a median value for ψ of roughly 0.1. Figure 2 plots the prior density of ψ.

Evaluating (21) numerically yields posterior means for β of 0.3487 ($s.e. = 0.1375$) and 0.4201 ($s.e. = 0.1556$) for math and reading achievement respectively. Figure 3 plots the posterior density of β for math and reading achievement and shows the upper and lower limits of minimum length 95 percent Bayesian credibility sets; these sets run from 0.0376 to 0.5824 for math achievement and from 0.0827 to 0.6976 for reading achievement. Combining the above prior on ψ with the Wald estimates of γ^2 given in Table 5 suggests that endogenous effects are a substantively important source of variation in academic achievement.

4.3 Discussion of estimates

The estimates of γ^2 reported in Table 5 suggest that social interactions substantively contributed to the learning process of Project STAR kindergarten students. There are two conceptually distinct

[31] Using the fact that for an exponential random variable $\int_{\psi=0}^{x} \pi(\psi|\lambda)\, d\psi = e^{-x\lambda}$, the appropriate choice for the scale parameter is $\lambda = \ln(0.001)/e \simeq 6.90776$.

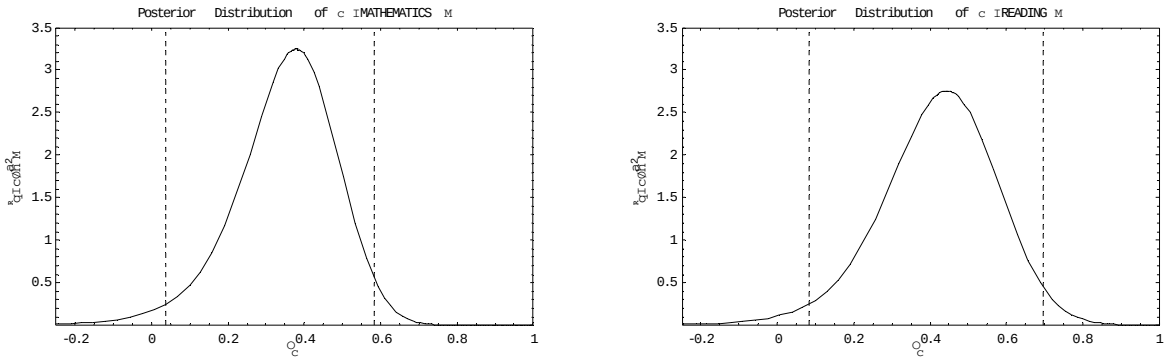

Figure 3: Posterior Densities for β
NOTES: Figure graphs numerically calculated posterior densities for β based on the Wald estimate of γ^2 given in Table 5 and an exponential prior on ψ with scale parameter, $\lambda \simeq 6.91$. The dashed vertical lines mark the 0.025 and 0.975 lower and upper probability tails of the densities (i.e., the lower and upper limits of a 95 percent credibility set for β).

ways to gauge the strength of the estimated effects. First, we can assess how changes in peer group composition affect the achievement of an individual student. Second, we can assess how changes in the assignment process of students to classrooms alter the overall distribution of achievement.

Table 7 reports the effects of hypothetical interventions of the first kind. As a benchmark row 1 reports the mean difference in normalized test scores between students at the 25^{th} and 75^{th} percentiles of the distribution of student ability, ε_{ci}, assuming normality and holding peer ability constant.[32] 'Above average' students score about one standard deviation higher on the math and reading tests than 'below average' students when placed in the same classroom.

Row 2 of the table compares test scores across students of equal ability but with above average peers versus below average peers ($\overline{\varepsilon}_c$ equal to 75^{th} and 25^{th} percentiles of the ε_{ci} distribution respectively). The effect of this hypothetical intervention is nearly as large as the direct effect of changes in own ability, consistent with a social multiplier near two.

Random assignment generated relatively homogenous Project STAR classrooms. Variations in peer ability as large as the above contrast are therefore not observed in the data. Since extrapolation beyond observed variation may lead to spurious conclusions, row 3 reports the effect of a change from the 25^{th} to 75^{th} percentiles of the *actual* distribution of peer composition across Project STAR classrooms. The effect of this intervention equals about 0.2 standard deviations for both math and reading, similar in magnitude to that of switching from a regular/regular-with-aide to a small classroom (row 5).

Row 4 assesses the effect of changes in teacher quality on achievement. Random matching

[32] A change from the 25^{th} to the 75^{th} percentiles approximately equals 2×0.67 standard deviations for a normally distribution random variable.

of students and teachers suggests that $\sigma_{\alpha\varepsilon} = \sigma_{\varepsilon\varepsilon} = 0$. The estimate of ς therefore captures excess variance solely due to heterogeneity in teacher quality, albeit amplified by the presence of endogenous social effects (i.e., $\varsigma = \sigma_{\alpha}^{2}/(1-\beta)^{2}$). Unfortunately ς is poorly identified by the Project STAR data. Its point estimate is slightly below zero with a large standard error for both math and reading achievement (see Table 5). I therefore take a very generous upper bound value for ς equal to its point estimate plus 1.96 standard deviations. The square root of this number is used as an estimate of the standard deviation of teacher quality. This upper bound estimate suggests that the reduced form effect of having an above average versus below average teacher is, at the very most, equal to a 0.4 standard deviation change in test scores.[33] Overall the effect of variation in peer group composition is, at the very least, similar in magnitude to those of commonly suggested strategies for raising student achievement such as improving teacher quality or reducing class size.

How would have deviations from random assignment altered the overall distribution of achievement amongst Project STAR kindergarten students? One feature of the linear-in-means model is that changes in the allocation of students across classrooms only affects the variance of achievement and not its overall mean. The model provides no traction on the equity versus efficiency trade-offs that emerge in the theoretical literature on sorting and peer group effects. Issues at the forefront of current debates about, for example, 'cream skimming' and the relative merits of different school choice policies (c.f., Nechyba 2003, Epple, Figlio and Romano 2004). This limitation notwithstanding, an assessment of the estimated values of γ^{2}'s implications for the relationship between student sorting across classrooms and inequality in achievement remains interesting.

Evidence reported in Vigdor and Nechbya (2004) implies an average within-classroom correlation of 'ability' – measured relative to a school-specific mean – of roughly 0.10 for 5^{th} graders attending North Carolina Public Schools. Clotfelter, Ladd and Vigdor (2004) attribute this correlation to 'teacher shopping' whereby parents of relatively advantaged children exert pressure on school administrators to ensure their children are assigned to superior teachers.[34]

The North Carolina data provides a natural benchmark for assessing the relationship between peer group effects, within-school sorting of students, and achievement inequality. How much would achievement inequality amongst Project STAR kindergarten students have increased if, instead of random assignment, the assignment mechanism mimicked that found in an average North Carolina school? To simplify, in what immediately follows I assume homoscedasticity of individual ability,

[33] This upper bound implies a standard deviation in teacher quality (in terms of test scores) larger than that found by Aaronson, Barrow, and Sander (2002) and almost three times as large as the effect size found by Rockoff (2003).

[34] The sample only includes schools with multiple 5^{th} grade classrooms. These rough estimates can be inferred from Table 1 of Vigdor and Nechyba (2004) by noting that the ratio of the within-*classroom* standard deviation in lagged test scores – 'ability' – to the within-*school* standard deviation approximately equals $\sqrt{1-\zeta_{\varepsilon\varepsilon}}$. Measuring 'ability' relative to the statewide mean, and hence including the effects of both within-school and cross-school sorting, increases the correlation coefficient to roughly 0.3.

$\sigma^2(w_c) \equiv \sigma^2$. The standard deviation in student achievement equals

$$\sigma_y(\zeta_{\varepsilon\varepsilon};\beta) = \sqrt{\sigma^2 + \varsigma + (\gamma^2 - 1)\,\sigma^2(\zeta_{\varepsilon\varepsilon} + (1 - \zeta_{\varepsilon\varepsilon})\,\mu_{1/M})} \tag{22}$$

where $\mu_{1/M} = E[1/M_c]$ and $\beta = (\varsigma,\ \gamma^2,\ \sigma^2,\ \mu_{1/M})$. With perfect sorting of students across classrooms, $\zeta_{\varepsilon\varepsilon} = 1$ and (22) collapses to $\sqrt{\gamma^2\sigma^2 + \varsigma}$. Perfect mixing requires that $\zeta_{\varepsilon\varepsilon} + (1 - \zeta_{\varepsilon\varepsilon}\mu_{1/M}) = 0$ and hence that $\zeta_{\varepsilon\varepsilon} = \mu_{1/M}/(\mu_{1/M} - 1)$ or that ability is negatively correlated across students within the same classroom. A simple measure of the increased achievement inequality associated with a shift from random assignment to modest sorting is $\sigma_y(\zeta_{\varepsilon\varepsilon};\beta)/\sigma_y(0;\beta)$, where $\zeta_{\varepsilon\varepsilon}$ measures the amount of within-classroom ability correlation in the counterfactual of interest.

Table 8 reports estimates of the above ratio for $\zeta_{\varepsilon\varepsilon}$ equal to 0.1, 0.3, and 1.0 respectively.[35] The first value corresponds to the modest level of within-school sorting suggested by Vigdor and Nechbya's (2004) data. The second and third values correspond to medium and perfect within-school sorting respectively. Shifting from random assignment to modest sorting suggests an increase in the standard deviation of math and reading achievement of about 9 and 12 percent respectively. A shift to perfect sorting suggests increases of 67 and 91 percent respectively.

To provide a benchmark with which to gauge the magnitude of these effects consider an intervention which eliminates black-white differences in 'ability' (i.e., background characteristics) for students within the same school. Decomposing individual heterogeneity into a systematic black-white difference and an idiosyncratic component we have

$$\varepsilon_{ci} = \text{BLACK}_{ci} \cdot \eta_b + \epsilon_{ci}, \qquad E[\epsilon_{ci}|\text{BLACK}_{ci}] = 0.$$

The unconditional variance of individual heterogeneity equals $\pi_b\eta_b^2 + \sigma_\epsilon^2$, where π_b is the proportion black for Project STAR kindergarten students. Eliminating black-white differences in background

[35] The parameters required to calculate the ratios reported in Table 8 are estimated by GMM using the moment vector

$$E\left[Z_c'\left(y_c - W_c\beta^*\right)\right] = 0 \tag{23}$$

where $\beta^* = \left(\beta',\ \eta_b,\ \eta_f\right)'$ and

$$y_c = \begin{pmatrix} g_c^b \\ M_c g_c^w \\ M_c^{-1} \\ \sum_{c=1}^{M_c} \widetilde{\text{BLACK}}_{ci}\widetilde{y}_{ci} \\ \sum_{c=1}^{M_c} \widetilde{\text{FREELUNCH}}_{ci}\widetilde{y}_{ci} \end{pmatrix}, \qquad W_c = \begin{pmatrix} 1 & g_c^w & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sum_{c=1}^{M_c} \widetilde{\text{BLACK}}_{ci}^{\,2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \sum_{c=1}^{M_c} \widetilde{\text{FREELUNCH}}_{ci}^{\,2} \end{pmatrix},$$

$$Z_c = \begin{pmatrix} 1 & q_c' & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sum_{c=1}^{M_c} \widetilde{\text{BLACK}}_{ci}^{\,2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \sum_{c=1}^{M_c} \widetilde{\text{FREELUNCH}}_{ci}^{\,2} \end{pmatrix}.$$

The large sample variance-covariance matrix is calculated in the usual way.

characteristics requires increasing ε_{ci} for all black students by $-\eta_b$. This lowers the standard deviation of achievement from

$$\sqrt{\sigma^2 + \varsigma + (\gamma^2 - 1)\,\sigma^2 \mu_{1/M}}$$

to

$$\sqrt{\left(\sigma^2 - \pi_b \eta_b^2\right) + \varsigma + (\gamma^2 - 1)\left(\sigma^2 - \pi_b \eta_b^2\right)\mu_{1/M}}.$$

Row 4 of Table 8 indicates that this intervention reduces achievement inequality by about 4 to 5 percent. An identical hypothetical intervention, but with respect to eligibility for free or reduced price school lunch, reduces the standard deviation in achievement by 7 to 8 percent.

Eliminating within-school black-white differences in background characteristics generates a large increase in average achievement (see columns 3 and 4 of Table 8). The effect on inequality, however, is modest since only a small reduction in individual, and hence peer group, heterogeneity occurs (i.e., $\pi_b \eta_b^2$ is small relative to σ^2). The direct effects of sorting are stronger. Sorting amplifies the effects of the large amount of 'naturally occurring' individual heterogeneity.[36]

5 Specification errors and testing

If the linear-in-means model correctly specifies how student ability, teacher quality and peer group composition collectively determine academic achievement, consistency follows directly from assumptions (17) and (18). Both of these assumptions are well-motivated by the experimental structure of Project STAR as well as consistent with several pieces of auxiliary evidence (see Tables 2, 3, and 4). The assumption that the educational production function give by (6) is correctly specified, however, is admittedly more tenuous.

5.1 Omnibus specification tests

One approach to assessing specification is to directly test the conditional moment restriction (19). Unfortunately, with q_c binary the unconditional moment restriction defining the Wald estimator contains all the information implied by (19) and no overidentifying restrictions are available for testing. With q_c non-binary the model is overidentified and, conditional on instrument validity, the extra restrictions can be used to test functional form assumptions (c.f., Angrist, Graddy and Imbens, 2000).

The class type indicator, q_c, exploits the exogenous variation in class size generated by Project STAR to form expected variance contrasts. The experimentally induced variation in class-size differs from the actual variation. Figure 4 shows that the distribution of class sizes across small and regular/regular-with-aide classrooms is partially overlapping. Furthermore the Wald estimator does

[36] Observe that the semielasticity of the standard deviation of achievement with respect to the intensity of within-school sorting is $\frac{\partial \ln \sigma_y}{\partial \zeta_{\varepsilon\varepsilon}}(\zeta_{\varepsilon\varepsilon};\beta) = \frac{1}{2}\frac{(\gamma^2-1)\sigma^2(1-\mu_{1/M})}{\sigma_y(\zeta_{\varepsilon\varepsilon};\beta)}$.

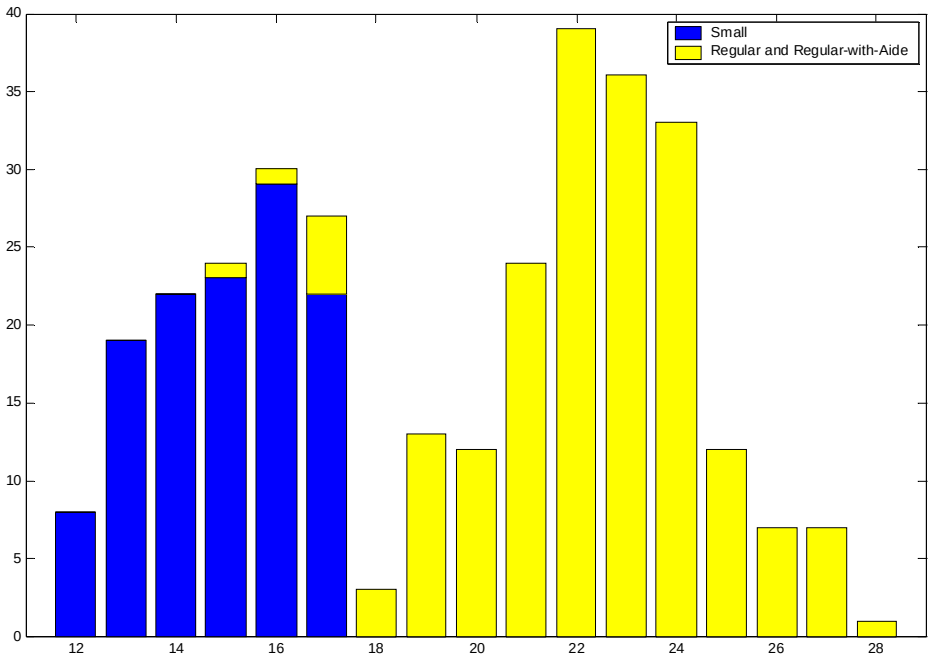

Figure 4: Distribution of Class Size Across Project STAR Kindergarten Classrooms
NOTES: The figure shows a histogram of the distribution of class-size across the 317 (out of 325) kindergarten classrooms included in the main sample (the structure of this sample is described in Section 2).

not use the substantial variation in size within class types. This is appropriate since the assignment mechanism producing teacher/size combinations *within class types* is not well documented. For example, it is possible the within randomly assigned class types, more senior teachers were systematically placed in smaller or larger classrooms.[37]

Notwithstanding the above concerns, if we are willing to treat the *actual* distribution of class size as exogenous, in the sense required for assumption (17) to hold, then we can base estimation on the moment $E\left[\rho\left(g_c,\theta\right)|M_c\right]=0$, which does imply overidentification. To implement this idea I partition the distribution of class size into three cells: small, with 12 to 16 students, medium, with 17 to 22 students, and large with 23 to 28 students.[38] This partition is entirely data-dependent and should not be confused with the experimental designation of classrooms as either small, regular or regular-with-aide.

[37] About one third of Project STAR classrooms are from schools with just three kindergartens. In these schools there would have been no scope for within-class-type discretion in teacher assignment.

[38] In principle, with M_c equalling one of only a finite number of values, one could estimate using a fully saturated set of moment restrictions. This does not seem sensible here due to the small sample size.

The three-bin partition generates a single overidentifying restriction as it allows estimates of γ^2 based on excess variance contrasts across both small and medium and medium and large classrooms. If the linear-in-means specification given by (6) is an adequate approximation to the true data generating process then the estimated γ^2 should be similar for both contrasts.

Table 9 reports three estimates of γ^2 for math and reading achievement. In column 1 $\gamma^2_{s/m}$ is estimated using contrasts across small- and medium-sized classrooms *only* ($N = 221$). For math achievement this contrast appears reasonably powerful with a first-stage F-statistic of 28.20, for reading achievement, however, identification is weak (F-statistic of 9.14). Column 2 uses contrasts across medium and large classrooms ($N = 214$). This contrast is too weak to generate meaningful estimates of $\gamma^2_{m/l}$, unsurprising since differences in $1/M_c$ are smaller across these two cells and it is variation in $1/M_c$ which drives differences in expected variance contrasts. Column 3 reports two-step GMM estimates of γ^2 using dummies for medium and large as instruments (with small being the excluded group). The point estimates of γ^2 are almost identical to those reported in Table 5, which only use experimentally-induced variation in class size. Row 3 of panels A and B report p-values of 0.2488 and 0.2828 for the Sargan-Hansen test of the restriction that $\gamma^2_{s/m} = \gamma^2_{m/l}$. While the test does not provide evidence against the null hypothesis of correct specification, its power to detect misspecification is presumably quite low in most directions.

5.2 Tests in specific directions: separability of teacher 'effectiveness' and class size

One (to this point) unstated implication of (17) is that class size and what I will now refer to as teacher 'effectiveness' (α_c) are separable. This is a rather strong assumption on the educational production function. While random assignment ensures that the underlying distribution of observed and unobserved teacher characteristics will be the same in small and large classrooms (given a sufficiently large sample), it seems plausible that the salience of specific teacher characteristics for student achievement may vary systematically with class size.[39] For example the importance of a teacher's ability to maintain 'order and discipline' may grow with class size. Similarly, a teacher's ability to customize his pedagogy to the needs of individual students may be more important in smaller classrooms.

Assume that teachers have L latent attributes $a_c = (a_{c1}, \ldots, a_{cL})'$. Random assignment ensures that

$$Var(a_c|q_c = 1) = Var(a_c|q_c) = \Xi_a.$$

The relative importance of each attribute for realized *teaching effectiveness* (α_c) however, may vary with class-size:

$$\alpha_c = a_c'\lambda_1 \cdot q_c + a_c'\lambda_0 \cdot (1 - q_c). \tag{24}$$

An implication of (24) is that the affect on student achievement from reducing class size will

[39] I thank Gary Chamberlain for clearly articulating this concern to me.

be teacher-specific. Under random assignment of teachers to class sizes, the conditional variance of teacher effectiveness will therefore differ across the two types of classrooms (even though the distribution of underlying latent teacher attributes will not). Formally

$$\sigma_{\alpha}^{2}(1) - \sigma_{\alpha}^{2}(0) = \lambda_1{}'\Xi_a\lambda_1 - \lambda_0{}'\Xi_a\lambda_0 \neq 0, \quad (\text{for } \lambda_1 \neq \lambda_0),$$

which violates (17). Without loss of generality we can normalize the latent teacher attributes such that $\Xi_a = I_L$. The relative effectiveness of teachers in small versus large classrooms *under random assignment to class type then equals*

$$\xi \overset{def}{=} \sqrt{\frac{\sigma_{\alpha}^{2}(1)}{\sigma_{\alpha}^{2}(0)}} = \frac{\|\lambda_1\|}{\|\lambda_0\|},$$

where $\|\cdot\|$ denotes the $l^2 - norm$. To interpret ξ it is useful to consider a simple thought experiment. Consider randomly drawing pairs of teachers from a common population, will reducing class size reduce ($\xi < 1$) or amplify ($\xi > 1$) the average difference in 'effectiveness' between the two teachers? Put differently, is variation in teacher effectiveness greater in small ($\xi > 1$) or large ($\xi < 1$) classrooms?

Relatively little is known about the educational production process. One view suggests that class size and some underlying notion of teacher 'ability' are complementary ($\xi < 1$). With complementarity we would expect that moving a common population of teachers to larger classrooms would, in addition to reducing *average* teacher effectiveness, increase its *variance.* In this case all teachers would perform relatively similarly in small classrooms with differences in teacher effectiveness only emerging in larger classrooms. Alternatively teacher 'ability' and class size could be substitutable ($\xi < 1$), with individual teacher characteristics being unimportant in large classrooms because, for example, anyone can effectively execute 'chalk and talk'. Finally it may be that while the salience of individual teacher attributes varies with class size, it is the case that on average – under random assignment to class type – the various effects wash out such that ξ is close to one.

The Wald-IV estimator will be inconsistent for γ^2 if $\xi \neq 1$, with a large sample bias of

$$\widehat{\gamma}^2 - \gamma^2 \xrightarrow{p} \frac{\left(\xi^2 - 1\right)\sigma_{\alpha}^{2}(0)}{E\left[g_c^w | q_c = 1\right] - E\left[g_c^w | q_c = 0\right]},$$

where for simplicity I have assumed that $Cov\left(\varepsilon_{ci}, \varepsilon_{cj}\right) = 0$ for $i \neq j$ and $Cov\left(a_l, \varepsilon_{ci}\right) = 0$ for $l = 1, \ldots, L$ (as seems reasonable for the Project STAR application). The bias will be downwards, or toward the no social interactions null, under complementarity and upwards under substitutability. Solving for ξ and replacing $\left(E\left[g_c^w | q_c = 1\right] - E\left[g_c^w | q_c = 0\right]\right)$ with its first stage estimate, denoted $\widehat{\phi}_2^w$,

we have approximately

$$\xi_1(\sigma_\alpha^2(0)\,|\widehat{\gamma}^2,\widehat{\phi}_2^w,\gamma^2)=1\simeq\sqrt{\gamma^2+\frac{\widehat{\phi}_2^w(\widehat{\gamma}^2-1)}{\sigma_\alpha^2(0)}}. \tag{25}$$

Assume that there are no social interactions ($\gamma^2=1$), (25) can be combined with assumptions on $\sigma_\alpha^2(0)$ to back out the degree of substitutability between teacher ability and class size that would be required to produce (large sample) estimates of the size found.

Table 10 reports the results of exercises of this type. To calibrate the experiments note that $\sigma_\alpha(0)$ equals the change in test scores associated with a one standard deviation change in teacher effectiveness *in regular and regular-with-aide classrooms.* The relevant distribution is the *within-school* distribution of teacher effectiveness, since the between-school variation in test scores has already been purged from the data. Rockoff (2004), using panel data methods, simple deconvolution procedures to deal with measurement error, and a sample of 'normal' sized classrooms from New Jersey, estimates $\sigma_\alpha(0)$ to be about 0.1. Aaronson, Barrow, and Sander's (2002) research, using Chicago Public Schools data, suggests a somewhat higher value for $\sigma_\alpha(0)$. A reasonable upper bound for $\sigma_\alpha(0)$ based on existing evidence is about 0.3, which is slightly above the value implied by the upper tail of the 95 percent confidence interval for $\widehat{\varsigma}$ (see Table 5).

For $\sigma_\alpha(0)=0.1$ the typical difference in effectiveness across a pair of teachers would have to be roughly 2.5 times larger in small versus large classrooms to produce γ^2 estimates of the size reported in Table 5, if in fact there were no peer effects. This is an implausibly large number. For $\sigma_\alpha=0.3$ the difference would have to 1.25 times larger, still quite a large effect. Overall identification appears to be strong enough to ensure a reasonable amount of robustness to bias caused by substitutability of teacher quality and class size.

A simple and direct test for substitutability/complementarity bias its to compare estimates of γ^2 based samples upon with large amounts of heterogeneity in teacher quality versus ones with little heterogeneity. If size and teacher quality are complementary then the estimate based on the first sample should be smaller than those based on the second. If teacher quality and class size are substitutes the opposite pattern will occur.

In the Project STAR dataset the only observed teacher covariate that is significantly related to test scores is years of teaching experience. I divide Project STAR schools (and hence classrooms) into two sets: in the first set the standard deviation of years teaching experience is greater than or equal to five, in the second set it is less than five. This partition is used to form subsamples with high and low degrees of heterogeneity in teacher quality. Table 11 reports separate estimates of γ^2 using these two subsamples.

The discussion emphasizes the math achievement results since those for reading achievement are not well identified, with first stage F-statistics all below 10. Column 1 reports the Wald estimate of γ^2 based on a comparison across small and large classrooms in schools with lots of heterogeneity

in years of teaching experience. Column 2 reports the estimate based on classrooms in schools with little experience heterogeneity. The two estimates are similar in magnitude, consistent with the null of separability.

Column 3 reports two-step GMM estimates of γ^2 using the entire sample with the small class type dummy and its interaction with a dummy for belonging to the high heterogeneity subsample ($Std\,(Exp_c) \geq 5$ years) serving as excluded instruments. Row 3 reports the p-value for a Sargan-Hansen test of the null that the high heterogeneity and low heterogeneity estimates of γ^2 are equal. There is little evidence of quantitatively important bias due to non-separability of teacher quality and class size in the educational production function.

The coefficient on the dummy variable for belonging to the high variance subsample is positive and statistically significant: schools with greater heterogeneity in experience also display greater between classroom variation in test scores. This result indirectly suggests that the comparison made in Table 11 should have real power to detect significant substitutability or complementarity bias.

5.3 Implications of heterogenous class-size effects

Project STAR provides strong evidence that lowering class size raises average achievement. *A priori*, however, there is no reason to assume that the effects of class size can be summarized by an additive homogenous 'treatment effect'. Plausible theories of educational production suggest that, among other possibilities, within-classroom variation in student achievement may be more or less dispersed, or more right- or left-skewed, in smaller than in larger classrooms. Such patterns might arise if the effects of class size are heterogeneous, as would occur, for example, if low ability students disproportionately benefit from a reduction in class size.[40]

The model used thus far has implicitly assumed a homogenous effect. A more flexible specification allows for heterogenous effects (c.f., Heckman and Vytlacil, 1998). Let 'realized' student ability, ε_{ci}, depend on 'latent' ability, ω_{ci}, and class type according to

$$\varepsilon_{ci} = \rho_0 \left(1 - q_c\right) \omega_{ci} + \rho_1 q_c \omega_{ci}.$$

If $\rho_1 > \rho_0$ a high ability student gains more from being placed in a small classroom than an average student, if $\rho_1 < \rho_0$ low ability students gain more. Henceforth ρ_0 is normalized to one without loss of generality. In the presence of heterogenous class size effects the difference in 'expected' between-group variance across small and large classrooms reflects two, potentially countervailing, forces. While class size variation induces a mechanical difference in the between-group variance of *latent* ability, ω_{ci}, variation in *realized* ability, ε_{ci}, is amplified or attenuated by heterogenous class size effects. For example, if low ability students benefit more from a reduction in

[40] Krueger and Whitmore (2001), using Project STAR data, report evidence that black students disproportionately benefit from class size reductions.

class size, then the dispersion of student achievement will be compressed in small relative to large classrooms. This compression of within-classroom student achievement also generates a mechanical reduction in the between-classroom variability of achievement.

In contrast, peer group effects, at least those of the linear-in-means form, do not alter the within-classroom variability of achievement. Their effects manifest themselves solely in the between-classroom variation of the data. An obvious concern is that the between-classroom variance contrasts observed in Project STAR reflect heterogenous class size effects, not social interactions.

To explore identification in the augmented model it is necessary to recast assumptions in terms of $\underline{\omega}_c$ and α_c. Assume that

$$E\left[\underline{\omega}_c, \alpha_c | q_c\right] = \left(0 \cdot \iota'_{M_c},\ \mu\left(q_c\right)\right)$$

and

$$Var\left(\underline{\omega}'_c, \alpha_c | q_c\right) = \begin{pmatrix} \sigma_\omega^2\left(1-\zeta\right) I_{M_c} + \zeta\sigma_\omega^2 \iota_{M_c}\iota'_{M_c} & \sigma_{\alpha\omega}\iota_{M_c} \\ \sigma_{\alpha\omega}\iota'_{M_c} & \sigma_\alpha^2 \end{pmatrix},$$

with ζ now equaling the within-classroom correlation of *latent student ability,* induced by any (within-class-type) sorting. Latent student ability is conditionally homoscedastic, a reasonable assumption given random assignment to class type. Realized student ability, however, is conditionally heteroscedastic since achievement gains associated with moving from a large to small classroom depend on a student's latent ability.

Under these assumptions the Wald-IV estimator of γ^2 has a large sample relative bias of

$$\frac{\widehat{\gamma}^2 - \gamma^2}{\gamma^2} \xrightarrow{p} \frac{\zeta}{1-\zeta} \frac{\rho_1^2 - 1}{\rho_1^2 E\left[M_c^{-1} | q_c = 1\right] - E\left[M_c^{-1} | q_c = 0\right]}, \tag{26}$$

which for $\zeta \neq 0$ and/or $\rho_1 \neq 1$ is different from zero. In the case of random assignment of students to classrooms (not just to class type) $\zeta = 0$, and the Wald-IV estimator remains consistent. This is because the mean difference in g_c^w across small versus large classrooms captures the net effects of the two effects of class size on 'expected' between-group variance in achievement. In the presence of sorting, however, this is not the case. Even if the within-classroom covariance of latent student ability in both small and large classrooms is the same, the covariance of realized student ability will not be; $E\left[g_c^w | q_c = 1\right] - E\left[g_c^w | q_c = 0\right]$ will no longer provide a consistent estimate of the difference in 'expected' between-group variance across small and large classrooms.

While the randomization tests reported in Section 2 provide no evidence of sorting on the basis of observed student characteristics, sorting on unobserved characteristics, and hence inconsistency, is a concern. Thirty-one of the 79 Project STAR schools had only three kindergarten classrooms. In these schools random assignment to *class type* is synonymous with random assignment to *classrooms.* In these schools with we can safely assume that $\zeta = 0$. The remaining 48 schools had up to 8 kindergarten classrooms and, consequently, scope for non-random (within-class-type) sorting

of students into classrooms (i.e., $\zeta \neq 0$). This may have occurred if school administrators sought to balance the composition of multiple classrooms of the same type ($\zeta < 0$) or if classrooms of the same type were purposively stratified by ability ($\zeta > 0$).

Here I use two facts to assess the degree of robustness to heterogenous class size effects: (a) the presence of heterogenous class size effects can be adduced from the within-classroom variation of the data alone, and (b) evidence on the nature of any non-random (within-class-type) sorting of students in larger schools – including that based on unobserved characteristics – can be found by comparing the magnitude of within-classroom variation in achievement across the two sets of schools.

Let $L_c = 1$ if the c^{th} classroom is located in one of the 48 large schools and zero otherwise. Making the auxiliary assumption that the variance of latent student ability is the same across both types of schools we have

$$E\left[M_c \cdot g_c^w | q_c, L_c\right] = \sigma_\omega^2 + \left(\rho_1^2 - 1\right)\sigma_\omega^2 \cdot q_c - \zeta_L \sigma_\omega^2 \cdot L_c - \left(\rho_1^2 - 1\right)\zeta_L \sigma_\omega^2 \cdot q_c \cdot L_c, \tag{27}$$

where ζ_L equals the within-classroom correlation in latent ability for students attending large schools.

Table 12 reports nonlinear least squares estimates of σ_ω, ρ_1, and ζ_L based on (27). For math achievement ρ_1 is significantly greater than 1. The point estimate suggests that a student with latent ability one-standard deviation above average would experience an achievement gain from switching to a small classroom 0.0639 standard deviations greater than that of the average student (the average effect equals 0.1631 standard deviations, see Table 15 below). For reading achievement there is no strong evidence of heterogenous class size effects.

The point estimates of ζ_L are negative but insignificantly different from zero for both math and reading achievement. These results provide little evidence of within-class-type sorting of students in larger schools. Any within-class-type sorting of students in larger schools was probably of the 'mixing' variety (c.f., Table 4). The point estimates for ρ_1 and ζ_L suggest an upward bias in estimates of γ^2 based on variance contrasts in *large schools alone* of 71 and 12 percent for math and reading achievement respectively (c.f., equation (26)). However the null of no bias is easily accepted in both cases.

On balance the estimates reported in Table 12 provide little evidence that heterogenous class size effects provide a compelling alternative rationalization for the variance contrasts observed in the Project STAR data. However, I am unable, given the limitations of the data in hand, to conclusively dispense with this possibility.

5.4 Outcome measurement error

The assumption that the outcome of interest is measured without error is untenable in most applications. Even if the variable in hand corresponds to the conceptual outcome of interest, it will

usually be only noisily observed by the econometrician. Even more often the available measure provides only a rough proxy for the actual outcome of interest. In the Project STAR application measurement error is of considerable concern. The available proxies for academic achievement, test scores in mathematics and reading, only loosely correspond to actual attainment in these subject areas. Reliability ratios for these tests are well below one, while even 'true' test scores would only provide imperfect measures of actual achievement.

Assume that we do not observe the true outcome of interest, y^*_{ci}, but only the noisy proxy $y_{ci} = y^*_{ci} + v_{ci}$, where $v_{ci}|M_c, q_c$ has constant conditional variance σ^2_v and is uncorrelated with both ε_{ci} and α_c. Although not required for the qualitative results, it is also convenient to assume that $\varepsilon_{ci}|M_c, q_c$ has constant conditional variance σ^2 as well. Under these conditions and assumptions (17) and (18) we have

$$\begin{aligned} E[g^b_c|M_c] &= \varsigma + \gamma^2\sigma^2 E[M_c^{-1}|q_c] + \sigma^2_v E[M_c^{-1}|q_c] \\ E[g^w_c|q_c] &= \sigma^2 E[M_c^{-1}|q_c] + \sigma^2_v E[M_c^{-1}|q_c] \end{aligned}$$

and a consequent downward bias in $\widehat{\gamma}^2$ of

$$\underset{N\to\infty}{plim}\ (\widehat{\gamma}^2 - \gamma^2) = \left(1-\gamma^2\right)(1-\kappa) < 0$$

where $\kappa = \sigma^2\left(\sigma^2 + \sigma^2_v\right)^{-1}$ is the signal-to-noise ratio in the test score data. Unlike ordinary least squares regression, dependent variable measurement error *does* result in a biased estimate of the coefficient of interest.

Occasionally auxiliary information on the nature and magnitude of dependent variable measurement error is available (e.g., from validation studies). In these cases it is straightforward to modify $\widehat{\gamma}^2$ to produce consistent estimates of γ^2. Given a known signal-noise-ratio, κ, a consistent estimate γ^2 and its asymptotic standard error is:

$$\widetilde{\gamma}^2_{EIV} = \widehat{\gamma}^2 \cdot \frac{1}{\kappa} - \left(\frac{1-\kappa}{\kappa}\right), \quad AVar\left(\widetilde{\gamma}^2_{EIV}\right) = \quad AVar\left(\widehat{\gamma}^2\right)/\kappa^2. \tag{28}$$

Table 13 uses the above formulae to adjust the γ^2 estimates reported in Table 5 for classical dependent variable measurement error of varying intensities. As the table makes clear, plausible levels for the signal-to-noise ratio between test scores and actual academic achievement, κ, can lead to substantial downward bias in $\widehat{\gamma}^2$.

If two (classically) noisy measures of the outcome of interest are available, with the measurement error across the two signals uncorrelated, it is possible to consistently estimate γ^2 as well as the signal-to-noise ratio for the two measures in one-step. Let

$$g^{b,(i,j)}_c = (\overline{y}^{(i)}_c - \mu^{(i)}(q_c))(\overline{y}^{(j)}_c - \mu^{(j)}(q_c)), \qquad g^{w,(i,j)}_c = \sum\nolimits_{i=1}^{M_c} \widetilde{y}^{(i)}_c \widetilde{y}^{(j)}_c$$

for noisy measurements $i, j = 1, 2$. The moment function

$$E\left[Z_c'\left(y_c - W_c\beta\right)\right] = 0, \tag{29}$$

where

$$y_c = \begin{pmatrix} g_c^{b,(1,2)} \\ M_c \cdot g_c^{w,(1,2)} \\ M_c \cdot g_c^{w,(1,1)} \\ M_c \cdot g_c^{w,(2,2)} \end{pmatrix}, \quad W_c = \begin{pmatrix} 1 & g_c^{w,(1,2)} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad Z_c = \begin{pmatrix} 1 & q_c & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

identifies $\beta = \left(\varsigma,\ \gamma^2,\ \sigma^2,\ \sigma^2 + \sigma_{v1}^2,\ \sigma^2 + \sigma_{v2}^2\right)'$.

Table 14 reports estimates of β based on the assumption that the math and reading test scores are independent signals of unobserved true achievement. Ideally independent test scores in the same subject area would be used for this exercise. Unfortunately multiple measures of this type are not included in the public release Project STAR data. The GMM estimate of γ^2 exceeds five, consistent with substantial measurement error bias. Indeed, the estimated signal-to-noise ratios for the math and reading tests are 0.68 and 0.65 respectively. These results illustrate the potential severity of the measurement error problem.

6 The excess sensitivity approach to identification

The most common and arguably current best practice test for social interactions is a reduced form test for excess sensitivity (e.g., Sacerdote 2001, Duncan *et al* 2003, Angrist and Lang 2004). This method exploits random assignment, or conditional random assignment, of individuals to groups to motivate simple least squares-based tests for social interactions. These tests are attractive since their plausibility is straightforward to evaluate and they are easy to implement. Graham and Hahn (2004) provide a formal overview of this approach.

Implementing these tests requires that in addition to outcomes, we observe a $K \times 1$ vector of individual-level characteristics, r_{ci}. This allows the individual heterogeneity term to be decomposed into observable and unobservable components, $\varepsilon_{ci} = r_{ci}'\eta + \epsilon_{ci}$. For ease of exposition and to more directly focus on the key differences between the excess sensitivity and variance approaches, consider the case where only endogenous social interactions are present (i.e., $\psi = 0$).

Substituting $\varepsilon_{ci} = r_{ci}'\eta + \epsilon_{ci}$ into (6) and rearranging to partition achievement into its within- and between-group components yields the reduced form

$$\begin{aligned} y_{ci} &= \overline{r}_c' \frac{\eta}{1-\beta} + \left(r_{ci} - \overline{r}_c\right)'\eta + u_{ci} \\ &= \overline{r}_c'\pi_b + \widetilde{r}_{ci}'\pi_b + u_{ci} \end{aligned} \tag{30}$$

where $u_{ci} = \frac{\alpha_c}{1-\beta} + \frac{1}{1-\beta}\overline{\epsilon}_c + (\epsilon_{ci} - \overline{\epsilon}_c)$.

Under random assignment of students to classrooms a least squares regression of y_{ci} on $\overline{r}_c$ and $\widetilde{r}_{ci}$ identifies $\pi = (\pi_b', \pi_w')' = \left(\left(\frac{1}{1-\beta}, 1\right) \otimes \eta'\right)'$. The null hypothesis of no social interactions can be assessed by testing the restriction $\pi_b = \pi_w$. Positive social interactions imply that $\pi_b > \pi_w$ or that there is *excess between-group sensitivity* in outcomes to between-group variation in characteristics. Note that π_b and π_w are identical to the coefficients in the within- and between-group regressions of y on r and hence, formally, the excess sensitivity test is a Hausman and Taylor (1981) test, although its motivation and interpretation are quite different (Graham and Hahn 2004).

In the augmented linear-in-means model (30) the full parameter vector is now $\theta = \left(\pi', \sigma_\epsilon^2, \sigma_\alpha^2, \gamma^2\right)'$, although only an estimate of π is required to implement the test.[41] The Wald statistic for the no social interactions null ($H_0 : \pi_{b0} - \pi_{w0} = 0$) is

$$W = N \cdot (\widehat{\pi}_b - \widehat{\pi}_w)' \left[\widehat{V}_{\pi_b} + \widehat{V}_{\pi_w}\right]^{-1} (\widehat{\pi}_b - \widehat{\pi}_w)$$

where $\widehat{V}_{\pi_b}$ and $\widehat{V}_{\pi_w}$ are $K \times K$ sub-matrices of the estimated large sample variance-covariance matrix for the reduced form excess sensitivity regression.[42] Given the simplifying assumption of no exogenous effects, H_0 is equivalent to the null that the social multiplier, γ, equals one since $\pi_b - \pi_w = (\gamma - 1)\eta$. In practice excess sensitivity is often assessed by running a regression of y_{ci} on r_{ci} and $\overline{r}_c$ and testing the joint significance of the latter set of coefficients. This test is numerically equivalent to the test for equality of the within- and between- reduced form coefficient vectors given above.

Table 16 implements the excess sensitivity test for social interactions using Project STAR math and reading test score data. The table reports estimates of π_b and π_w (full regression results are reported in Table 15). Included in r_{ci} are the four individual characteristics mentioned in Section 2: dummies for student race, gender, and free lunch eligibility as well as an age measure. Also included in the regression are basic classroom-level controls and a school fixed effect (since it is only the within-school variation in observable classroom composition that is arguably idiosyncratic). Note that these estimates require random assignment to classrooms, not just to class type.

To facilitate comparisons the between- and within-group coefficients are reported side-by-side in columns 1 and 2 with column 3 giving the difference. Under positive social interactions the magnitude of the between-group coefficients (in absolute value) should be greater than the corresponding within-group coefficients. The omnibus test of for no excess sensitivity is easily accepted with p-values of 0.3117 and 0.1670 for math and reading test scores respectively. The only individual-level covariate displaying significant excess sensitivity is gender.

Overall the excess sensitivity tests provide little evidence of peer group effects. However, they

[41] Observe that σ^2, the variance of individual heterogeneity term in the model without covariates, equals $\eta'\Sigma_{rr}\eta + \sigma_\epsilon^2$, where $E\left[r_{ci}' r_{ci}\right] = \Sigma_{rr}$ and σ_ϵ^2 is the variance of the residual $\varepsilon_{ci} - E^*\left[\varepsilon_{ci} | r_{ci}\right]$.

[42] Typically the variance-covariance matrix for $\widehat{\pi}$ is of the Huber-White variety with 'clustering' at the level of social groups.

also provide little evidence against the existence of even quite large effects. Table 16 also reports tests for the restriction $\pi_b = 2 \cdot \pi_w$, which would hold if the true social multiplier were two, a large value and in excess of that implied by both the math and reading estimates of γ^2 reported in Table 5. The test accepts with a p-value of 0.3608 for math achievement and marginally accepts with a p-value of 0.0770 for reading achievement. The excess sensitivity regressions are consistent with both very small and very large levels of peer group effects. In principle, the idiosyncratic variation in observable class composition generated by the Project STAR experiment provides an ideal opportunity to implement an excess sensitivity test for social interactions; unfortunately the test is uninformative.

7 Relative power of excess variance and excess sensitivity tests

An obvious reason why the excess sensitivity test to fails reject the no social interactions null is low power. The idiosyncratic variation in classroom composition generated by random assignment of students and teachers to classrooms, while exogenous, may be too small to reliably detect the presence of social interactions. This section derives and compares the large sample power functions for both the excess variance and excess sensitivity tests. The power functions provide insight into what design features and test combinations will reliably detect social interactions when present. The analysis demonstrates that the excess variance test provides a substantially more powerful test for social interactions than the conventional excess sensitivity test for designs 'like' Project STAR.

In order to derive an interpretable expressions, I make the auxiliary assumptions of homoscedasticity and joint normality of $(\underline{\varepsilon}_c', \alpha_c)$ throughout. I also assume, without loss of generality, that $\sigma_{\varepsilon\varepsilon} = \sigma_{\alpha\varepsilon} = 0$.[43] These assumptions are made *ex post* in order to derive interpretable expressions. That is, I assume that normality and homoscedasticity happen to occur in the population being sampled from but that these additional restrictions are not used in estimation, although in principal they could be used to derive a more efficient estimator.[44] Finally, again for simplicity, I continue to assume the absence of any exogenous social effects.

7.1 Excess variance test power

Before deriving the actual power function, some additional notation and results are required. First write the excess variance model in simultaneous equations form with the structural equation

$$g_c^b = \varsigma + \gamma^2 g_c^w + u_c \tag{31}$$

and associated first stage

$$g_c^w = \phi_1^w + q_c' \phi_2^w + v_c. \tag{32}$$

[43] The normality assumption is only required for the excess variance power function.

[44] In particular, these two additional assumptions could be used to derive a maximum likelihood estimator.

Under assumptions (17) and (18) we have $E[u_c|q_c] = E[u_c|q_c] = 0$ and hence $\phi_2^w = E[g_c^w|q_c = 1] - E[g_c^w|q_c = 0]$, the denominator of the population Wald estimator. Also write the reduced form regression as

$$g_c^b = \phi_1^b + q_c'\phi_2^b + e_c \tag{33}$$

with $\phi_1^b = \varsigma + \gamma^2\phi_1^w$ and $\phi_2^b = \gamma^2\phi_2^w$.

The power of the excess variance test depends in part on the strength of the first stage. A unitless measure of instrument strength is the *concentration parameter* $\kappa_0 = N\phi_2^{w\prime}V_w^{-1}\phi_2^w$, where V_w is the large sample variance of $\widehat{\phi}_2^w$ (c.f., Bound, Jaeger and Baker 1995, Staiger and Stock 1997). The larger the expected variance contrast, relative to the precision with which it is estimated, the stronger the instrument. Under the normality assumption it is straightforward to show that

$$\phi_2^w = \sigma^2\left(E\left[M_c^{-1}|q_c = 1\right] - E\left[M_c^{-1}|q_c = 0\right]\right),$$

where the large sample distribution of $\widehat{\phi}_2^w$ is

$$\sqrt{N}(\widehat{\phi}_2^w - \phi_2^w) \xrightarrow{D} \mathcal{N}(0, V_w)$$

with

$$V_w = 2\sigma^4\left\{\frac{1}{\pi}E\left[\frac{1}{M_c^2}\frac{1}{M_c - 1}|q_c = 1\right] + \frac{1}{1-\pi}E\left[\frac{1}{M_c^2}\frac{1}{M_c - 1}|q_c = 0\right]\right\},$$

where $\pi = E[q_c]$ or the fraction of small classrooms. The concentration parameter is therefore

$$\kappa_0 = \frac{N}{2}\frac{\left(E\left[M_c^{-1}|q_c = 1\right] - E\left[M_c^{-1}|q_c = 0\right]\right)^2}{\pi^{-1}E\left[M_c^{-2}(M_c - 1)^{-1}|q_c = 1\right] + (1-\pi)^{-1}E\left[M_c^{-2}(M_c - 1)^{-1}|q_c = 0\right]}. \tag{34}$$

One interesting and unique feature of (34) is that it depends *only* on the marginal distribution of class size and not on any features of the distribution of ε_{ci}, or individual ability. This is a direct consequence of the auxiliary assumptions of normality and homoscedasticity, which ensure that the square of the second moment of ε_{ci} is proportional to its fourth moment. Although these two assumptions are unlikely to hold in practice, at least exactly, (34) makes clear the importance of group size variation in ensuring strong identification and is is also likely to be helpful for researchers interested in designing sampling plans and experiments for detecting social interactions. Equation (34) also enters the power function for the excess variance test.

We now derive the power function for the no social interactions null

$$H_0 : \gamma^2 = 1. \tag{35}$$

The standard difficulty in deriving an asymptotic power function is that if the null is false any consistent test will reject with probably one as $N \to \infty$. Here I follow standard practice and avoid

degenerate ⊤-shaped power functions by evaluating limiting power for a fixed null under a sequence of alternative data generating processes (DGPs). In these alternatives the social interactions parameter, γ_N^2, follows a 'Pitman drift'. In particular γ_N^2 evolves with N according to $\gamma_N^2 = 1+\delta_0/\sqrt{N}$ for some δ_0. While the alternative DGP approaches the hypothesized null as the sample size grows, the rate of shrinkage is parameterized to ensure that the difference $\sqrt{N}\left(\gamma_N^2 - \gamma_0^2\right)$ remains fixed at δ_0. The nuisance parameters σ^2 and σ_α^2 remain fixed across the entire sequence of DGPs.

Under these conditions it is straightforward to show that the scaled difference$\sqrt{N}\left(\widehat{\gamma}_N^2 - 1\right)$ converges in distribution to a normal random variable with mean δ_0 and variance equal to $\Sigma_{\gamma\gamma}(\theta_0)$, where $\Sigma_{\gamma\gamma}(\theta_0)$ is the lower right-hand element of the asymptotic variance-covariance matrix for $\widehat{\theta} = \left(\widehat{\varsigma}, \widehat{\gamma}^2\right)$ evaluated at the null. The Wald statistic for (35) therefore converges in distribution to a non-central $\chi^2_{1,\lambda}$ random variable with non-centrality parameter $\lambda = \delta_0^2 / \Sigma_{\gamma\gamma}(\theta_0)$. Since, under Pitman drift, $\sqrt{N}\left(\widehat{\gamma}_N^2 - 1\right) \to \left(\gamma_N^2 - 1\right) \equiv \delta_0$ this suggests that we can approximate the finite sample behavior of the Wald statistic for the null of no social interactions for a given DGP in the sequence with a non-central $\chi^2_{1,\lambda}$ distribution with non-centrality parameter

$$\lambda = \left(\gamma_N^2 - 1\right)\left(\frac{V_{b0}}{N\left(\phi_2^w\right)^2} + \frac{1}{\kappa_0}\right)^{-1}\left(\gamma_N^2 - 1\right), \tag{36}$$

where $\left(\frac{V_b}{N(\phi_2^w)^2} + \frac{1}{\kappa_0}\right)^{-1}$ equals $\left(\Sigma_{\gamma\gamma}(\theta_0)/N\right)^{-1}$ as shown in Appendix B.

The second term in the middle parentheses of (36) captures the asymptotic penalty which arises from having to estimate the first-stage coefficient, ϕ_2^w, or the difference in 'expected' between-group outcome variance across the two sets of groups defined by the binary instrument q_c. The size of this penalty is inversely proportional to the concentration parameter for instrument strength derived above, κ_0, and would equal zero if ϕ_2^w were known. For ϕ_2^w unknown, strong instruments lead to a more powerful test. Recall that κ_0 depends only on the marginal distribution of group size.

The first term in the middle parentheses equals the infeasible Wald statistic for the null hypothesis of no difference in the actual between-group variance in outcomes across the two sets of groups defined by q_c. Appendix B gives the requisite components required to form a complete analytical expression for $V_{b0}/N\left(\phi_2^w\right)^2$. The leading term of $V_{b0}/N\left(\phi_2^w\right)^2$, however, captures the main factors driving power. Under the no social interactions null this term equals

$$\frac{1}{N}\frac{2}{\pi(1-\pi)}\left(\frac{\sigma_\alpha^2}{\sigma^2}\right)^2\frac{1}{\left(E\left[M_c^{-1}|q_c=1\right] - E\left[M_c^{-1}|q_c=0\right]\right)^2}$$

and power therefore increases in sample size, N; increases in $\pi(1-\pi)$, which is maximized at $\pi = 1/2$ or when the subsamples defined by q_c are of the same size; decreases in σ_α^2/σ^2, the ratio of variances for group- and individual-level heterogeneity; and increases in $E\left[M_c^{-1}|q_c=1\right] - E\left[M_c^{-1}|q_c=0\right]$, highlighting the importance of group-size variation.

To summarize, power depends on the precision with the first stage and reduced form regressions are estimated, or equivalently the precision of the numerator and denominator of $\widehat{\gamma}^2_{WALD}$. The precision of the first stage is unaffected by the degree of unobserved group-level heterogeneity as well as the magnitude of any social interactions since it is based on the within-group variation of the data (which is purged of these influences). The precision of the reduced form, in contrast, does depend on the amount of group-level heterogeneity in the population as well as on the marginal distribution of group-size.

7.2 Excess sensitivity test power

As in the excess variance case we evaluate the power to reject the no excess sensitivity restriction under a sequence of local alternative DGPs where γ_N, the social multiplier, evolves with N such that

$$H_1^N : \gamma_N = 1 + \delta_0/\sqrt{N}.$$

Observe that $\pi_{bN} - \pi_{wN} = \delta_0\eta/\sqrt{N}$, which approaches zero as the sample size grows; the alternative DGP thus remains in a $1/\sqrt{N}$ neighborhood of the no social interactions null. Under this sequence of alternatives, H_1^N, we can show, proceeding analogously to the excess variance case, that the Wald statistic for no excess sensitivity converges to a non-central $\chi^2_{K,\lambda}$ random variable with non-centrality parameter,

$$\lambda = \eta'\delta_0 \left[V_{\pi_b}(\theta_0) + V_{\pi_w}(\theta_0)\right]^{-1} \delta_0\eta. \tag{37}$$

Since $\gamma_N - \gamma_0 = \delta_0/\sqrt{N}$ by construction, (37) implies that we can approximate the distribution of the Wald statistic for a given DGP in the sequence of alternatives by a $\chi^2_{K,\lambda}$ distribution with non-centrality parameter,

$$\lambda = N \cdot (\gamma - 1)\, \eta' \left[V_{\pi_b}(\theta_0) + V_{\pi_w}(\theta_0)\right]^{-1} \eta\, (\gamma - 1)\,. \tag{38}$$

Appendix B shows that $V_{\pi_b}(\theta_0) + V_{\pi_w}(\theta_0)$, evaluated at the no social interactions/excess sensitivity null ($\gamma_0 = 1$), equals

$$\left(\sigma^2_\epsilon + \mu_M\sigma^2_\alpha + \frac{\sigma^2_\epsilon}{\mu_M - 1}\right) \cdot \Sigma^{-1}_{rr},$$

where $\mu_M = E[M_c]$ and $E[r'_{ci}r_{ci}|M_c] = E[r'_{ci}r_{ci}] = \Sigma_{rr}$. Substituting into (38) and rearranging yields

$$\lambda = N \cdot (\gamma - 1)^2 \frac{\eta'\Sigma_{rr}\eta\sigma^2_\alpha}{\mu_M + \frac{\sigma^2_\epsilon}{\sigma^2_\alpha}\frac{1}{\mu_M - 1}}. \tag{39}$$

Equation (39) illustrates the main factors which determine the large sample power of excess sensitivity tests. First, power is increasing in the size of the social multiplier. Second it is increasing in the ratio of *observed* variation in individual-level ability, $\eta'\Sigma_{rr}\eta$, to *unobserved* classroom-level

variation in teacher effectiveness, σ^2_α. A weakness of the excess sensitivity test is that, unlike the excess variance one, it only exploits variation contained in observed individual characteristics.

Power is decreasing in mean group size, μ_M, at a rate which depends on the ratio $\sigma^2_\epsilon/\sigma^2_\alpha$. Increases in group size have two offsetting effects on the power of the excess sensitivity test. While larger group sizes improve the precision with which the within-group reduced form coefficients, π_w, are estimated, they reduce the precision with which the corresponding between-group coefficients are estimated, π_b. The second effect dominates for large enough μ_M. As $\mu_M \to \infty$ the power of the excess sensitivity test approaches zero. Excess sensitivity tests will be most powerful when they are based on variation in a salient observed characteristic (large η), with large marginal variances (Σ_{rr}), when groups are of modest size (low μ_M), and are similar in terms of unobserved environmental factors (small σ^2_α).

While these results are intuitive and *ex post* unsurprising, they are not sufficiently internalized in much of the recent wave of empirical social interactions research. While random assignment to groups *does* help to identify social interactions, it *does not* guarantee reliable detection of them in small samples when groups are moderately large, as is the case in Project STAR.

7.3 Calibrated power comparisons

This section calibrates the excess variance and excess sensitivity power functions to the Project STAR dataset. The variance-component parameters σ^2 and σ^2_α as well as the social multiplier, γ, are estimated by maximum likelihood under the assumption that $\underline{y}_c|M_c$ is multivariate normal.[45] Full results are given in Table 17. Estimates of η are taken from the within-group reduced form coefficients reported in Table 15, and Σ_{rr} is replaced by its sample analog. Panel B of Table 18 lists the parameter values used for calibration.

Panel A uses (36) and (39) to compute the power of the excess sensitivity and variance tests to detect social interactions across repeated samples drawn from the calibrated population (designed to mimic the Project STAR dataset). For math achievement, given a true social multiplier of 1.75, the excess sensitivity and variance tests correctly detect social interactions about 85 and 99 percent of the time respectively. The relative power results for reading achievement are qualitatively similar with rejection rates of 73 and 98 percent. These are substantial power differences; while the excess variance test reliably rejects the no social interactions null, the excess sensitivity test fails to reject 15 to 25 percent of the time for the level of social interactions found in the Project STAR data. The odds of correct rejection using the excess variance test are 26 and 19 times greater for math and reading achievement respectively.

Panel A also computes the inner and outer inverse power functions for the two tests using the methods of Andrews (1989). The inner inverse equals the value of the social multiplier below which the given test *fails* to reject at least 50 percent of the time. In samples where $\gamma < \gamma^{II}$, the given

[45] Homoscedasticity of ε_{ci} is also assumed such that $\sigma^2(M_c) = \sigma^2$; although this is not required for identification.

test will be worse than one based on a coin flip. The outer inverse equals the value of the social multiplier above which the test rejects at least 95 percent of the time; for $\gamma \geq \gamma^{OI}$ the test will reliably reject the no social interactions null. Values of γ between the two thresholds define a region of the alternative where the test, while having some power, is unreliable.

Overall the exercise confirms the superior power of the excess variance test in the Project STAR design. Unfortunately both tests lack power to detect small to modest levels of social interactions. In the case of math achievement, for example, the excess sensitivity and variance tests are very unreliable when the true social multiplier is below 1.5 and 1.4 respectively. These calculations suggests that the methods and datasets typically employed by researchers may not be able to reliably detect modest levels of social interactions and that greater emphasis should be placed on confidence intervals.

8 Conclusion and areas for further research

This paper has outlined a new approach to identifying peer group effects based on excess variance contrasts across groups of differing sizes. The method exploits excess variance intuition and is robust to the presence of confounding group-level heterogeneity. Applying the method to the Project STAR dataset suggests social multipliers between 1.07 and 2.31 and 1.05 to 3.07 for math and reading achievement respectively. The estimates provide strong evidence that peer group composition was a substantively important input into the learning process of Project STAR students. The results of a battery of specification tests, as well as detailed considerations of rival models, suggests that the social interaction interpretation of these results is appropriate.

By virtue of random assignment, excess sensitivity tests also plausibly identify social interactions in the Project STAR data. These tests, however, provide no evidence of peer effects. This apparent contradiction has a straightforward explanation: for designs like Project STAR, tests based on excess variance contrasts are substantially more powerful than conventional excess sensitivity tests. While this result is design specific, it does underscore a more general concern that the datasets and methods typically used by researchers may not reliably detect substantively important levels of social interactions if present. For the Project STAR design, even the excess variance test is worse than a coin flip for values of the social multiplier below 1.4, equivalent to a reaction function slope, β, of about 0.3. Most researchers would consider this a large peer group effect. This suggests that empirical social interactions research should pay careful attention to test power and make greater use of confidence intervals.

If peer effects are important for learning, the design of educational policies needs to reflect this information. Low peer group quality, even at a young age, may have long-run consequences. Currie and Thomas (1999), working with data from the British National Child Development Study, find that student performance as early as age seven affects subsequent educational attainment, employment and adult earnings conditional on controlling for a large number of observable characteristics.

Krueger (2003), in a brief literature review, suggests that existing evidence is consistent with the belief that a one standard deviation increase in elementary school test scores is associated with about 8 percent higher earnings as an adult. This conclusion is also consistent, at least qualitatively, with the evidence that parents are willing to pay substantial housing price premiums to locate in the catchment areas of high performing schools (Black 1999). For these reasons producing estimates of the size of peer group effects on academic achievement across different grade levels that are both credibly and strongly identified remains a high priority for future research.

This paper has worked within the linear-in-means framework. This model is the workhorse of applied social interactions research and is a natural point of departure for exploring identification through conditional covariance restrictions. Even this simple model has resulted in a problematic empirical literature and a robust coherent body of research based on it has yet to fully emerge (c.f., Manski 1993, Durlauf 2002).

Unfortunately, as noted earlier, the linear-in-means model is unable to answer some of the most important questions raised by theoretical peer group effects models. The linearity of the model implies that any redistribution of peers will leave mean achievement unaffected, with only inequality in achievement changing. The model provides no traction on the equity versus efficiency trade-offs that typically emerge in the theoretical models and dominate public discussions. Clearly identification and estimation of richer models of peer group effects is a natural next step.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. (2002). "Teachers and student achievement in the Chicago Public High Schools," *Federal Reserve Bank of Chicago WP 2002-28.*

Akerlof, George A. (1997). "Social distance and social decisions," *Econometrica* 65 (5): 1005 - 1027.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis.* New York: John Wiley & Sons.

Andrews, Donald W.K. (1989). "Power in econometric applications," *Econometrica* 57 (5): 1059 - 1090.

Angrist, Joshua D., Kathryn Graddy and Guido W. Imbens. (2000). "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish," *Review of Economics Studies* 67 (3): 499 - 527.

Angrist, Joshua D. and Kevin Lang. (2004). "Does school integration generate peer effects? Evidence from Boston's Metco program," *American Economic Review* 94 (5): 1613 - 1634.

Bayer, Patrick, Fernando Vendramel Ferreira, and Robert McMillan. (2004). "Tiebout sorting, social multipliers and the demand for school quality," *Mimeo, Yale University.*

Becker, Gary S. and Kevin M. Murphy. (2000). *Social Economics: Market Behavior in a Social Environment.* Cambridge, MA: Harvard University Press.

Black, Sandra E. (1999). "Do better schools matter? Parental valuation of elementary education," *Quarterly Journal of Economics* 114 (2): 577 - 599.

Boozer, Michael A. and Stephen E. Cacciola. (2004). "Inside the 'black box' of Project STAR: estimation of peer effects using experimental data," *Mimeo.*

Bound, John, David A. Jaeger and Regina M. Baker. (1995). "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak," *Journal of the American Statistical Association* 90 (430): 443 - 450.

Brock, William A. and Steven N. Durlauf. (2001). "Interactions-based Models," *Handbook of Econometrics 5*: 3297 - 3380 (J. Heckman & E. Leamer, Eds.). Amsterdam: North-Holland.

Clotfelter, Charles T., Helen F. Ladd and Jacob L. Vigdor. (2004). "Teacher sorting, teacher shopping, and the assessment of teacher effectiveness," *Mimeo.*

Cooper, Russell and Andrew John. (1988). "Coordinating coordination failures in Keynesian models," *Quarterly Journal of Economics* 103 (3): 441 - 463.

Cressie, Noel and Timothy R.C. Read. (1984). "Multinomial goodness-of-fit tests," *Journal of the Royal Statistical Society B* 46 (3): 440 - 464.

Currie, Janet and Duncan Thomas. (1999). "Early test scores, socioeconomic status and future outcomes," *NBER Working Papers No. 6943.*

Duncan, Greg J. *et al.* (2003). "Empathy or antipathy? The consequence of racially and socially diverse peers on attitudes and behaviors," *Mimeo.*

Durlauf, Steven N. (2002). "Groups, social influences and inequality: a memberships theory perspective on poverty traps," *SSRI Working Paper 2002-18.*

Durlauf, Steven N. and Marcel Fafchamps. (2004). "Social capital," *SSRI Working Paper 2004-12* (forthcoming *Handbook of Economic Growth*).

Epple, Dennis, David Figlio and Richard Romano. (2004). "Competition between private and public schools: testing stratification and pricing restrictions," *Journal of Public Economics* 88 (7-8): 1215 - 1245.

Epple, Dennis and Holger Sieg. (1999). "Estimating equilibrium model of local jurisdictions," *Journal of Political Economy* 107 (4): 645 - 681.

Finn, Jeremy D., Susan B. Gerber, Charles M. Achilles and Jayne Boyd-Zaharias. (2001). "The enduring effects of small classes," *Teachers College Record* 103 (2): 145 - 183.

Gaviria, Alejandro. (2000). "Increasing returns and the evolution of violent crime: the case of Colombia," *Journal of Development Economics* 61 (1): 1 - 25.

Glaeser, Edward L., Bruce Sacerdote and José A. Scheinkman. (1996). "Crime and social interactions," *Quarterly Journal of Economics* 111 (2): 507 - 548.

Glaeser, Edward L., Bruce Sacerdote and José A. Scheinkman. (2003). "The Social Multiplier," *Journal of the European Economic Association* 1 (2-3): 345 - 353.

Glaeser, Edward and José A. Scheinkman. (2001). "Measuring social interactions," *Social Dynamics: 83 - 130* (S. Durlauf & P. Young). Cambridge, MA: MIT Press.

Glaeser, Edward and José A. Scheinkman. (2003). "Nonmarket interactions," *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress 1*: 339 - 369 (M. Dewatripont *el al*, Eds.). Cambridge: Cambridge University Press.

Gould, Eric D., Victory Lavy and M. Daniele Paserman. (2004). "Long-term classroom peer effects: evidence from random variation in the enrollment of immigrants," *Mimeo.*

Graham, Bryan S. (2005). "Small sample properties of GMM and GEL estimation and inference procedures for social interaction models," *Mimeo, Harvard University.*

Graham, Bryan S. and Jinyong Hahn. (2004). "Identification and estimation of the linear-in-means model of social interactions," *Mimeo, Harvard University.*

Hanushek, Eric A. (1971). "Teacher characteristics and gains in student achievement: estimation using micro data," *American Economic Review* 60 (2): 280 - 288.

Hausman, Jerry A. and William E. Taylor (1981). "Panel data and unobservable individual effects," *Econometrica* 49 (6): 1377 - 1398.

Heckman, James and Edward Vytlacil. (1998). "Instrumental variable methods for the correlated random coefficient model: estimating the average return to schooling when the return is correlated with schooling," *Journal of Human Resources* 33 (4): 974 - 987.

Hoxby, Caroline M. (2002). "The power of peers: how does the makeup of a classroom influence achievement?" *Education Next* 2 (2): 57 - 63.

Krueger, Alan B. (1999). "Experimental estimates of education production functions," *Quarterly Journal of Economics* 114 (2): 497 - 532.

Krueger, Alan B. (2003). "Economic considerations and class size," *Economic Journal* 113 (485): F34 - F63.

Krueger, Alan B. and Diane M. Whitmore. (2001). "Would smaller classes help close the black-white achievement gap?" *Mimeo, Princeton University*

Lazear, Edward P. (2001). "Educational production," *Quarterly Journal of Economics* 116 (3): 777 - 803.

Lee, Valerie E. and David T. Burkam. (2002). *Inequality at the Starting Gate: Social Background Differences as Children Begin School.* Washington D.C.: Economic Policy Institute.

Lefgren, Lars. (2004). "Educational peer effects and the Chicago public schools," *Journal of Urban Economics* 56 (2): 169 - 191.

MaCurdy, Thomas E. (1982). "The use of time series processes to model the error structure of earnings in a longitudinal data analysis," *Journal of Econometrics* 18 (1): 83 - 114.

Manski, Charles F. (1993). "Identification of endogenous social effects: the reflection problem," *Review of Economic Studies* 60 (3): 531 - 542.

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

Moffitt, Robert A. (2001). "Policy interventions, low-level equilibria and social interactions," *Social Dynamics*: 45 - 82 (S. Durlauf & P. Young, Eds.). Cambridge, MA: MIT Press.

Nechyba, Thomas J. (2003). "Introducing school choice into multi-district public school systems," *The Economics of School Choice*: 145 - 194 (Caroline Hoxby, Ed.). Chicago: University of Chicago Press

Newey, Whitney K. (1985). "Generalized method of moments specification testing," *Journal of Econometrics* 29 (3): 229 - 256.

Newey, Whitney K., Joaquim J.S. Ramalho and Richard J. Smith. (2005). "Asymptotic bias for GMM and GEL estimators with estimated nuisance parameters," *Identification and Inference for Econometric Models: A Festschrift in Honor of Thomas Rothenberg*: 245 - 281 (D.W.K Andrews & J.H. Stock, Eds.). Cambridge: Cambridge University Press.

Newey, Whitney K. and Richard J. Smith. (2004). "Higher order properties of GMM and generalized empirical likelihood estimators," *Econometrica* 72 (1): 219 - 255.

Owen, Art. B. (2001). *Empirical Likelihood.* New York: Chapman & Hall/CRC.

Piketty, Thomas. (2000). "Theories of persistent inequality and intergenerational mobility," *Handbook of Income Distribution 1*: 430 - 476 (A. Atkinson & F. Bourguignon). Amsterdam: North Holland.

Rigobon, Robert. (2004). "Identification through heteroscedasticity," *Review of Economics and Statistics* 85 (4): 777 - 792.

Rockoff, Jonah (2004). "The impact of individual teachers on student achievement: evidence from panel data," *American Economic Review* 94 (2): 247 - 252.

Rubin, Donald B. (1981). "Estimation in parallel randomized experiments," *Journal of Educational Statistics* 6 (4): 377 - 400.

Sacerdote, Bruce. (2001). "Peer effects with random assignment: results for Dartmouth roommates," *Quarterly Journal of Economics* 116 (2): 681 - 704.

Solon, Gary, Marianne E. Page and Greg J. Duncan. (2000). "Correlations between neighboring children in their subsequent educational attainment," *Review of Economics and Statistics* 82 (3): 383 - 392.

Staiger, Douglas and James H. Stock. (1997). "Instrumental variables regression with weak instruments," *Econometrica* 65 (3): 557 - 586.

Stock, James H., Jonathan H. Wright and Motohiro Yogo. (2002). "A survey of weak instruments and weak identification in generalized method of moments," *Journal of Business and Economic Statistics* 20 (4): 518 - 529.

Stock, James H. and Motohiro Yogo. (2005). "Testing for weak instruments in linear IV regression," *Identification and Inference for Econometric Models: A Festschrift in Honor of Thomas Rothenberg*: 80 - 108 (D.W.K Andrews & J.H. Stock, Eds.). Cambridge: Cambridge University Press.

Topa, Giorgio. (2001). "Social interactions, local spillovers and unemployment," *Review of Economic Studies* 68 (2): 261 - 295.

A Forms for g_c^b and g_c^w required when only a random subsample of individuals in each group is observed

Let M_c^* denote the actual number of individuals sampled in the c^{th} group, with $M_c \geq M_c^*$continuing to denote group size. Let $\overline{y}_c^*$ denote the mean outcome across sampled group members, with $\overline{y}_c$ denoting the true (unobserved) group mean. Redefine g_c^b and g_c^w to equal

$$\begin{aligned} g_c^b &= \left(\overline{y}_c^* - \mu_y\left(q_c\right)\right)^2 - \left(\frac{1}{M_c^*} - \frac{1}{M_c}\right)\frac{1}{M_c^* - 1}\sum_{i=1}^{M_c^*}\left(y_{ci} - \overline{y}_c^*\right)^2 \\ g_c^w &= \frac{1}{M_c}\frac{1}{M_c^* - 1}\sum_{i=1}^{M_c^*}\left(y_{ci} - \overline{y}_c^*\right)^2. \end{aligned}$$

Observe that g_c^w and g_c^b continue to have the conditional expectations

$$\begin{aligned} E\left[g_c^w|q_c\right] &= E\left[\frac{\sigma^2\left(w_c\right)\left(1 - \zeta_{\varepsilon\varepsilon}\left(w_c\right)\right)}{M_c}|q_c\right] \\ E\left[g_c^b|q_c\right] &= \varsigma\left(q_c\right) + \gamma^2 E\left[\frac{\sigma^2\left(w_c\right)\left(1 - \zeta_{\varepsilon\varepsilon}\left(w_c\right)\right)}{M_c}|q_c\right], \end{aligned}$$

and hence all the estimators discussed in the main text remain the same subject to redefinition of g_c^w and g_c^b.

B Asymptotic power of excess variance and sensitivity tests for social interactions

B.1 Power function for excess variance test

This appendix derives the large sample power function for the excess variance test of the null of no social interactions given by (36) above. Recall that homoscedasticity and normality are assumed to hold.

The following preliminary results will proved useful in the derivation. Let x equal the matrix of group sizes $x = (M_c, \ldots, M_N)'$ and let $y_c^+ = (y_{c1} - \overline{y}_c, \ldots, y_{cM_c} - \overline{y}_c, \overline{y}_c)'$. The conditional variance of the $(M_c + 1) \times 1$ vector y_c^+ is

$$\Omega_c^+\left(x\right) = \begin{pmatrix} \frac{M_c - 1}{M_c}\sigma^2 & -\frac{\sigma^2}{M_c} & \cdots & -\frac{\sigma^2}{M_c} & 0 \\ -\frac{\sigma^2}{M_c} & \frac{M_c - 1}{M_c}\sigma^2 & & \vdots & \vdots \\ \vdots & & \ddots & -\frac{\sigma^2}{M_c} & \vdots \\ -\frac{\sigma^2}{M_c} & & -\frac{\sigma^2}{M_c^*} & \frac{M_c - 1}{M_c}\sigma^2 & 0 \\ 0 & \cdots & \cdots & 0 & \varsigma + \gamma^2\frac{\sigma^2}{M_c} \end{pmatrix},$$

From Anderson (1984, p. 49) we can show that under normality (suppressing the c subscript)

$$E\left[y_i^{+2}y_k^{+2}\right] = \omega_{ii}\omega_{kk} + 2\omega_{ik}\omega_{ik}, \qquad E\left[y_i^{+4}\right] = 3\omega_{ii}^2$$

and hence that

$$\begin{aligned} Var\left(g_c^w|x\right) &= \frac{2\sigma^4}{M_c^2}\frac{1}{M_c-1}, \\ Var\left(g_c^b|x\right) &= 2\left[\varsigma + \gamma^2\frac{\sigma^2}{M_c}\right]^2, \\ Cov\left(g_c^w, g_c^b|x\right) &= 0. \end{aligned}$$

Now consider the reduced form multivariate regression of $\left(g_c^b, g_c^w\right)$ on $\mathbf{1}\left(q_c=1\right)$ and $\mathbf{1}\left(q_c=0\right)$ *with no constant* and where $\mathbf{1}\left(\cdot\right)$ denotes the indicator function. Using results on the fourth centered moments given above we can show that the variance-covariance matrix for the coefficient vector on the between-group part of this regression will be

$$V_b^* = 2\sigma^4\gamma^4 diag\left\{\frac{1}{\pi}E\left[\left(\lambda_\alpha + \frac{1}{M_c}\right)^2|q_c=1\right],\ \frac{1}{1-\pi}E\left[\left(\lambda_\alpha + \frac{1}{M_c}\right)^2|q_c=0\right]\right\} \tag{40}$$

where $\lambda_\alpha = \frac{\sigma_\alpha^2}{(1-\beta)^2}\frac{1}{\sigma^2\gamma^2}$. The variance for the within-group part will be

$$V_w^* = 2\sigma^4 diag\left\{\frac{1}{\pi}E\left[\frac{1}{M_c^2}\frac{1}{M_c-1}|q_c=1\right],\ \frac{1}{1-\pi}E\left[\frac{1}{M_c^2}\frac{1}{M_c-1}|q_c=0\right]\right\}. \tag{41}$$

We can relate these terms to the standard 'first stage' and reduced form regressions of g_c^w and g_c^b on 1 and q_c respectively by noting that $V_w = V_{w1}^* + V_{w0}^*$ and $V_b = V_{b1}^* + V_{b0}^*$, where the subscripts denote the elements of V_b^* and V_w^* corresponding to the $q_c=1$ and $q_c=0$ cases. Denote the four reduced form coefficients from this regression as ϕ_1^{*b}, ϕ_0^{*b}, ϕ_1^{*w} and ϕ_0^{*w}; the difference of the first and second coefficients in the two regressions equals the corresponding coefficients on q_c in the first stage and reduced form regressions, i.e., $\phi_2^w = (\phi_1^{*w} - \phi_0^{*w})$ and $\phi_2^b = \left(\phi_1^{*b} - \phi_0^{*b}\right)$.

An identifying moment corresponding to the Wald estimate of γ^2 is

$$\psi_c\left(z_c,\theta\right) = \begin{pmatrix} \mathbf{1}\left(q_c=1\right) \\ \mathbf{1}\left(q_c=0\right) \end{pmatrix}\left(g_c^b - \varsigma - \gamma^2 g_c^w\right),$$

where q_c is a dummy variable for small class/group size.

We have $\Gamma_0 = E\left[\partial\psi_N\left(\theta_0\right)/\partial\theta'\right]$ equal to

$$\Gamma_0 = -\begin{pmatrix} \pi & \pi\sigma^2 E\left[M_c^{-1}|q_c=1\right] \\ 1-\pi & \left(1-\pi\right)\sigma^2 E\left[M_c^{-1}|q_c=0\right] \end{pmatrix}$$

where $\pi = E[q_c]$. Again using results on the multivariate normal we can show that $\Lambda_0 = E\left[\psi_c(\theta_0)\,\psi_c(\theta_0)'\right]$ is the 2×2 diagonal matrix

$$\Lambda_0 = 2\sigma^4\gamma^4 diag\left\{\pi E\left[a\left(M_c;\rho_\alpha\right)|q_c=1\right],\ \ (1-\pi)\,E\left[a\left(M_c;\rho_\alpha\right)|q_c=0\right]\right\},$$

where $a(M_c;\lambda_\alpha)$ equals

$$\begin{aligned} a(M_c;\lambda_\alpha) &= \frac{1}{2\sigma^4\gamma^4}\left\{Var(g_c^b|x)+\gamma^4 Var(g_c^b|x)+2\gamma^2 Cov(g_c^w,g_c^b|x)\right\} \\ &= \left(\lambda_\alpha+\frac{1}{M_c}\right)^2+\frac{1}{M_c^2}\frac{1}{M_c-1}. \end{aligned}$$

Standard GMM results yield a large sample variance-covariance matrix of $\left(\Gamma_0'\Lambda_0^{-1}\Gamma_0\right)^{-1}$. Multiplying $\Gamma_0'\Lambda_0^{-1}\Gamma_0$ out yields:

$$\begin{aligned} &\Gamma_0'\Lambda_0^{-1}\Gamma_0 \\ &= \frac{1}{2\sigma^4\gamma^4}\begin{pmatrix} \frac{1}{\frac{1}{\pi}E[a(M_c;\lambda_\alpha)|q_c=1]}+\frac{1}{\frac{1}{1-\pi}E[a(M_c;\lambda_\alpha)|q_c=0]} & \frac{\sigma^2E\left[M_c^{-1}|q_c=1\right]}{\frac{1}{\pi}E[a(M_c;\lambda_\alpha)|q_c=1]}+\frac{\sigma^2E\left[M_c^{-1}|q_c=0\right]}{\frac{1}{1-\pi}E[a(M_c;\lambda_\alpha)|q_c=0]} \\ \frac{\sigma^2E\left[M_c^{-1}|q_c=1\right]}{\frac{1}{\pi}E[a(M_c;\lambda_\alpha)|q_c=1]}+\frac{\sigma^2E\left[M_c^{-1}|q_c=0\right]}{\frac{1}{1-\pi}E[a(M_c;\lambda_\alpha)|q_c=0]} & \frac{\sigma^4E\left[M_c^{-1}|q_c=1\right]^2}{\frac{1}{\pi}E[a(M_c;\lambda_\alpha)|q_c=1]}+\frac{\sigma^4E\left[M_c^{-1}|q_c=0\right]^2}{\frac{1}{1-\pi}E[a(M_c;\lambda_\alpha)|q_c=0]} \end{pmatrix}. \end{aligned}$$

Using standard results on partitioned inverse the asymptotic variance-covariance matrix for $\widehat{\gamma}^2_{WALD}$ is therefore[46]

$$AVar\left(\widehat{\gamma}^2_{WALD}\right) = \left(\Gamma_0'\Lambda_0^{-1}\Gamma_0\right)^{-1}/N = \frac{\gamma^4}{N}\begin{Bmatrix} \frac{\sigma^4E\left[M_c^{-1}|q_c=1\right]^2}{\frac{2\sigma^4}{\pi}E[a(M_c;\lambda_\alpha)|q_c=1]}+\frac{\sigma^4E\left[M_c^{-1}|q_c=0\right]^2}{\frac{2\sigma^4}{1-\pi}E[a(M_c;\lambda_\alpha)|q_c=0]} \\ -\frac{\left(\frac{\sigma^2E\left[M_c^{-1}|q_c=1\right]}{\frac{2\sigma^4}{\pi}E[a(M_c;\lambda_\alpha)|q_c=1]}+\frac{\sigma^2E\left[M_c^{-1}|q_c=0\right]}{\frac{2\sigma^4}{1-\pi}E[a(M_c;\lambda_\alpha)|q_c=0]}\right)^2}{\frac{1}{\frac{2\sigma^4}{\pi}E[a(M_c;\lambda_\alpha)|q_c=1]}+\frac{1}{\frac{2\sigma^4}{1-\pi}E[a(M_c;\lambda_\alpha)|q_c=0]}} \end{Bmatrix}^{-1}.$$

From (40) and (41) we can rewrite $\frac{2\sigma^4}{\pi}E\left[a\left(M_c;\lambda_\alpha\right)|q_c=1\right] = \frac{1}{\gamma^4}V_{b1}^* + V_{w1}^*$ and similarly for the

[46] When only M_c^* outcomes (out of M_c) are observed in the c^{th} group the asymptotic variance of the $\widehat{\gamma}^2_{WALD}$ remains the same with $a(M_c, M_c^*;\lambda_\alpha)$ replacing $a(M_c;\lambda_\alpha)$ and defined to equal

$$a(M_c,M_c^*;\lambda_\alpha) = \frac{1}{2\sigma^4\gamma^4}\left\{Var(g_c^b|x)+\gamma^4 Var(g_c^b|x)+2\gamma^2 Cov(g_c^w,g_c^b|x)\right\},$$

where

$$\begin{aligned} Var(g_c^w|x) &= \frac{2\sigma^4}{M_c^2}\frac{1}{M_c^*-1}, \\ Var(g_c^b|x) &= 2\gamma^4\sigma^4\left[\lambda_\alpha+\frac{1}{M_c}\right]^2-2\gamma^2\sigma^4\left(\frac{1}{M_c^*}-\frac{1}{M_c}\right)^2\left[\lambda_\alpha+\frac{1}{M_c}\right]+\sigma^4\left(\frac{1}{M_c^*}-\frac{1}{M_c}\right)^2\left[\frac{M_c^*+1}{M_c^*-1}\right], \\ Cov(g_c^w,g_c^b|x) &= -\frac{2\sigma^4}{M_c^*}\left(\frac{1}{M_c^*}-\frac{1}{M_c}\right)\left(\frac{1}{M_c^*-1}\right). \end{aligned}$$

$q_c = 0$ case. Substituting we get

$$\begin{aligned} AVar\left(\widehat{\gamma}^2_{WALD}\right) &= \frac{\gamma^4}{N}\left\{\frac{(\phi_1^{*w})^2}{\frac{1}{\gamma^4}V_{b1}^* + V_{w1}^*} + \frac{(\phi_0^{*w})^2}{\frac{1}{\gamma^4}V_{b0}^* + V_{w0}^*} - \frac{\left(\frac{\phi_1^{*w}}{\frac{1}{\gamma^4}V_{b1}^*+V_{w1}^*} + \frac{\phi_0^{*w}}{\frac{1}{\gamma^4}V_{b0}^*+V_{w0}^*}\right)^2}{\frac{1}{\frac{1}{\gamma^4}V_{b1}^*+V_{w1}^*} + \frac{1}{\frac{1}{\gamma^4}V_{b0}^*+V_{w0}^*}}\right\}^{-1} \\ &= \frac{\gamma^4}{N}\left\{\frac{(\phi_1^{*w})^2 + (\phi_0^{*w})^2 - 2\phi_1^{*w}\phi_0^{*w}}{\frac{1}{\gamma^4}V_{b1}^* + V_{w1}^* + \frac{1}{\gamma^4}V_{b0}^* + V_{w0}^*}\right\}^{-1} \\ &= \frac{\gamma^4}{N}\left\{\frac{\frac{1}{\gamma^4}V_b + V_w}{(\varphi_2^w)^2}\right\} \\ &= \gamma^4\left\{\frac{V_b}{N(\gamma^2\varphi_2^w)^2} + \frac{1}{\kappa_0}\right\}. \end{aligned}$$

B.2 Power function for excess sensitivity test

The large sample variance covariance matrix for the ordinary least squares estimator of $\pi = (\pi_b', \pi_w')' = \left(\left(\frac{1}{1-\beta}, 1\right) \otimes \eta'\right)'$ is

$$V_\pi(\theta_0) = \sigma_\epsilon^2 E\left[w_c' w_c\right]^{-1} \cdot E\left[w_c' \Omega_c(\theta_0) w_c\right] \cdot E\left[w_c' w_c\right]^{-1},$$

where $w_{ci} = (\widetilde{r}_{ci}', \overline{r}_c')'$, $w_c = (w_{c1}, \ldots, w_{cM})'$, and $\theta = \left(\pi', \sigma_\epsilon^2, \sigma_\alpha^2, \gamma^2\right)'$. Observe that σ^2, the variance of individual heterogeneity term in the model without covariates, equals $\eta'\Sigma_{rr}\eta + \sigma_\epsilon^2$, where $E[r_{ci}' r_{ci}] = \Sigma_{rr}$ and σ_ϵ^2 is the *residual* variance of $\varepsilon_{ci} - E^*[\varepsilon_{ci}|r_{ci}]$. Under the auxiliary assumptions of homoscedasticity and the presence of only endogenous social interactions we have

$$\Omega_c(\theta_0) = I_{M_c} + \left[\rho_{\alpha,\epsilon}\gamma^2 + \left(\gamma^2 - 1\right)\frac{1}{M_c}\right]\iota_{M_c}\iota_{M_c}',$$

where $\rho_{\alpha,\epsilon} = \sigma_\alpha^2/\sigma_\epsilon^2$ and $\gamma^2 = 1/(1-\beta)^2$.

The $E[w_c' w_c]$ terms evaluates to

$$E\left[w_c' w_c\right] = E\begin{pmatrix} \sum_{i=1}^{M_c} \widetilde{r}_{ci}'\widetilde{r}_{ci} & 0 \\ 0 & M_c\overline{r}_c'\overline{r}_c \end{pmatrix} = \begin{pmatrix} \mu_M - 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \Sigma_{rr},$$

where $\mu_M = E[M_c]$ and the equality $E[r_{ci}' r_{ci}|M_c] = E[r_{ci}' r_{ci}]$ has been assumed.

The $E\left[w_c'\Omega_c w_c\right]$ term simplifies as follows:

$$\begin{aligned} E\left[w_c'\Omega_c w_c\right] &= E\left[\sum_{i=1}^{M_c} w_{ci}'w_{ci} + \left(\rho_{\alpha,\epsilon}\gamma^2 + \frac{\gamma^2-1}{M_c}\right)\sum_{i=1}^{M_c}\sum_{j=1}^{M_c} w_{ci}'w_{cj}\right] \\ &= \begin{pmatrix} \mu_M - 1 & 0 \\ 0 & \gamma^2\left(1+\mu_M\rho_{\alpha,\epsilon}\right) \end{pmatrix} \otimes \Sigma_{rr}, \end{aligned}$$

where use has been made of iterated expectations.

Combing terms the large sample variance-covariance matrix divided by N for $\widehat{\pi}$ is

$$AVar\left(\widehat{\pi}\right) = \gamma^2\sigma^2\left[\begin{pmatrix} \frac{1}{\gamma^2}\frac{1}{\mu_M-1} & 0 \\ 0 & 1+\mu_M\rho_{\alpha,\epsilon} \end{pmatrix} \otimes \Sigma_{rr}^{-1}/N\right].$$

Using $AVar\left(\widehat{\pi}\right)$, the asymptotic variance associated with the difference, $\widehat{\pi}_b - \widehat{\pi}_w$, again divided by N, is thus:

$$AVar\left(\widehat{\pi}_b - \widehat{\pi}_w\right) = \gamma^2\sigma^2\left(1+\mu_M\rho_{\alpha,\epsilon} + \frac{1}{\gamma^2}\frac{1}{\mu_M-1}\right)\cdot\Sigma_{rr}^{-1}/N,$$

which can be rearranged to give the expression in the text.

Table 1: Means and Standard Deviations for Individual- and Classroom-Level Project STAR Kindergarten Variables

	(1) Sample Mean	(2) Standard Deviation	(3) Within-School r.m.s.e	(4) Within-Class r.m.s.e
Individual Test Scores				
Math Scores	0	1	0.8935	0.8393
Reading Scores	0	1	0.8954	0.8496
Individual Level Variables				
Black	0.3291	0.4699	0.2380	0.2390
Girl	0.4854	0.4998	0.4996	0.4995
FreeLunch	0.4825	0.4997	0.4189	0.4170
Date-of-Birth (DOB)	0.1116	0.3513	0.3482	0.3484
Classroom Level Variables				
$\overline{\text{Black}}_c$	0.3198	0.4071	0.0476	–
$\overline{\text{Girl}}_c$	0.4858	0.1184	0.1178	–
$\overline{\text{FreeLunch}}_c$	0.4774	0.2889	0.1072	–
$\overline{\text{DOB}}_c$	0.1104	0.0927	0.0791	–

Notes: Reported statistics in column 1 of the 'Individual Test Scores' and 'Individual Level Variables' panels are individual variable means for the 6,172 (out of 6,325) kindergarten students included in the regression analysis. These students are from 317 (out of 325) classrooms in the 79 Project STAR schools. Omitted students are from 8 classrooms that either had missing teacher data or could not be clearly disaggregated into separate classrooms. Test score statistics are based on observations from 5,724 and 5,646 students for the mathematics and reading exams respectively. Reported test scores have been normalized by the mean and standard deviation of the distribution of all scores. Column 2 reports the standard deviation of these variables. Columns 3 and 4 report the (residual) root mean squared error (r.m.s.e) associated with a regression of each variable on a vector of school and classroom dummy variables respectively. The 'Classroom Level Variables' panel reports means and overall and standard deviations for the 317 classroom averages of each of the individual characteristics and the r.m.s.e associated with the regression of these averages on a vector of schools dummies.

Table 2: Conditional Variance Contrasts Across Small and Large Project STAR Classrooms

	(1) $Var(\overline{x}_c\|SMALL)$	(2) $Var(\overline{x}_c\|REG)$	(3) $Var(\overline{x}_c\|SMALL) - Var(\overline{x}_c\|REG)$	(4) $\frac{Var(\overline{x}_c\|SMALL)}{Var(\overline{x}_c\|REG)}$
Test Scores				
MATH	0.1672	0.0946	0.0726	1.7676
	(0.0231)	(0.0110)	$(0.0256)^{**}$	$(0.3197)^{*}$
READING	0.1623	0.0860	0.0762	1.8861
	(0.0313)	(0.0120)	$(0.0335)^{*}$	$(0.4485)^{*}$
Individual-Level				
BLACK	0.0020	0.0015	0.0005	1.3465
	(0.0005)	(0.0003)	(0.0006)	(0.4401)
GIRL	0.0120	0.0066	0.0054	1.8136
	(0.0023)	(0.0008)	$(0.0024)^{*}$	$(0.4089)^{*}$
FREELUNCH	0.0148	0.0079	0.0069	1.8692
	(0.0013)	(0.0008)	$(0.0015)^{**}$	$(0.2538)^{**}$
DOB	0.0059	0.0040	0.0019	1.4747
	(0.0008)	(0.0005)	$(0.0009)^{*}$	$(0.2588)^{+}$
Group-Level				
MASTERS	0.1979	0.1360	0.0619	1.4551
	(0.0193)	(0.0107)	$(0.0221)^{**}$	$(0.1825)^{**}$
BLACKTEACHER	0.0801	0.0564	0.0237	1.4206
	(0.0153)	(0.0089)	(0.0177)	(0.3519)
EXPERIENCE	29.0593	25.5142	3.5450	1.1389
	(3.8104)	(2.4279)	(4.5182)	(0.1845)
CLAD	2.3077	1.8130	0.4948	1.2729
	(0.3928)	(0.2673)	(0.4752)	(0.2867)

Notes: Columns 1 and 2 report estimates of the conditional between-group variance for each of the listed individual and teacher variables by class type. Columns 3 and 4 report the difference and ratio of these variances respectively. 'Data' are class means of residuals from a preliminary regression of each of the listed variables on a matrix of school dummies and the class-type indicator variable. Squares of these between-classroom residuals are then regressed on the indicator variables SMALL and REGULAR (with no constant) to produce the estimates given in columns 1 and 2. The Huber-White variance-covariance matrix associated with this regression provides asymptotically valid standard errors in this case, even with the two-step procedure. Standard errors are reported in parentheses. The career ladder variances are computed using the 289 classrooms/teachers (number small = 112) with non-missing values for this variable. All results based on the full dataset described in the text. '**', '*', and '+' indicate that the difference of column 1 and 2 variances is statistically different from zero (or their ratio different from one) at the 1, 5 and 10 percent level respectively.

Table 3: Covariance of Group- and Individual-Level Characteristics in Small and Large Classrooms

Conditional Covariance Contrasts	(1) $Cov\left(\overline{x}_c, z_c \mid SMALL\right)$	(2) $Cov\left(\overline{x}_c, z_c \mid REG\right)$	(3) $Cov\left(\overline{x}_c, z_c \mid SMALL\right)$ $-Cov\left(\overline{x}_c, z_c \mid REG\right)$
Black ×			
Masters	−0.0004	0.0021	−0.0026
	(0.0021)	(0.0014)	(0.0025)
BlackTeacher	0.0011	0.0008	0.0002
	(0.0008)	(0.0005)	(0.0010)
Experience	0.0136	0.0130	0.0006
	(0.0224)	(0.0146)	(0.0267)
CLAD	0.0017	−0.0052	0.0070
	(0.0049)	(0.0037)	(0.0061)
Girl ×			
Masters	−0.0019	0.0012	−0.0030
	(0.0050)	(0.0024)	(0.0056)
BlackTeacher	0.0030	−0.0021	0.0051
	(0.0030)	(0.0014)	(0.0033)
Experience	0.0730	0.0623	0.0107
	(0.0535)	(0.0318)	(0.0623)
CLAD	−0.0180	−0.0125	−0.0055
	(0.0185)	(0.0096)	(0.0208)
FreeLunch ×			
Masters	0.0002	0.0011	−0.0008
	(0.0050)	(0.0022)	(0.0054)
BlackTeacher	0.0002	−0.0031	0.0033
	(0.0024)	(0.0013)	(0.0027)
Experience	−0.0677	−0.0116	−0.0561
	(0.0462)	(0.0264)	(0.0532)
CLAD	0.0048	0.0021	0.0027
	(0.0140)	(0.0067)	(0.0156)
DOB ×			
Masters	0.0021	0.0027	−0.0006
	(0.0031)	(0.0015)	(0.0035)
BlackTeacher	0.0010	0.0008	0.0002
	(0.0019)	(0.0008)	(0.0021)
Experience	0.0268	−0.0181	0.0449
	(0.0280)	(0.0220)	(0.0356)
CLAD	−0.0055	−0.0046	−0.0009
	(0.0088)	(0.0068)	(0.0111)

Notes: Columns 1 and 2 report estimates of the conditional between-group covariance for each of the listed individual and teacher variable pairs by class size, while column 3 reports the difference in these covariances. Cross products of the between-classroom residuals described in the notes to Table 2 are regressed on the indicator variables SMALL and REGULAR (with no constant) to produce the covariance estimates given in columns 1 and 2. All results based on the full dataset described in the text. '**', '*', and '+' indicate that the difference of column 1 and 2 covariances is statistically different from zero at the 1, 5 and 10 percent level respectively.

Table 4: Within-Classroom Covariance of Individual-Level Characteristics in Small and Large Classrooms

Within-Classroom Covariance Contrasts	(1) $Cov\,(x_{ci,}x_{cj,}\vert SMALL)$	(2) $Cov\,(x_{ci,}x_{cj,}\vert REG)$	(3) $Cov\,(x_{ci,}x_{cj,}\vert SMALL)$ $-Cov\,(x_{ci,}x_{cj,}\vert REG)$
Black	−0.00143	−0.0010	−0.0005
	(0.0006)	(0.0003)	(0.0006)
Girl	−0.0020	−0.0035	0.0015
	(0.0014)	(0.0008)	(0.0016)
FreeLunch	−0.0006	−0.0009	0.0003
	(0.0022)	(0.0008)	(0.0024)
DOB	−0.0021	−0.0016	−0.0006
	(0.0008)	(0.0004)	(0.0009)

Notes: Raw data are residuals from the preliminary regression described in the notes to Table 2. The product of these residuals times their leave-own-out mean are then regressed on the indicator variables SMALL and REGULAR (with no constant) to produce the covariance estimates given in columns 1 and 2. Standard errors are clustered at the classroom level. Column 3 reports the difference of the column 1 and 2 estimates. '**', '*', and '+' indicate that the difference of column 1 and 2 covariances is statistically different from zero at the 1, 5 and 10 percent level respectively.

Table 5: Wald Estimates of γ^2 for Normalized Math and Reading Achievement Test Scores

WALD-IV ESTIMATES		(1) Small (q_c= 1)	(2) Large (q_c= 0)	(3) Small - Large
Panel A : Math Achievement				
OBSERVED BETWEEN VARIANCE	$\overline{g^b}$	0.1626	0.0922	0.0704
		(0.0229)	(0.0110)	(0.0254)
EXPECTED BETWEEN VARIANCE	$\overline{g^w}$	0.0531	0.0303	0.0228
		(0.0030)	(0.0011)	(0.0032)
WALD ESTIMATE	$\widehat{\gamma}^2_{WALD}$	–	–	3.0891
				(1.0357)
	$\widehat{\varsigma}_{WALD}$	–	–	−0.0156
				(0.0381)
$F_{(df1,df2)}$ $1^{st}-Stage$		–	$51.01_{(1,316)}$	$p = 0.0000$
Panel B : Reading Achievement				
OBSERVED BETWEEN VARIANCE	$\overline{g^b}$	0.1533	0.0824	0.0708
		(0.0301)	(0.0119)	(0.0324)
EXPECTED BETWEEN VARIANCE	$\overline{g^w}$	0.0511	0.0330	0.0182
		(0.0041)	(0.0019)	(0.0045)
WALD ESTIMATE	$\widehat{\gamma}^2_{WALD}$	–	–	3.8967
				(1.8294)
	$\widehat{\varsigma}_{WALD}$	–	–	−0.0460
				(0.0668)
$F_{(df1,df2)}$ $1^{st}-Stage$		–	$16.27_{(1,316)}$	$p = 0.0001$
NUMBER OF CLASSROOMS	N	123	194	317

Notes: Estimates based on the full sample of 6,172 (out of 6,325) students across 317 (out of 325) classrooms described in the text. Row 1 of Panel A reports an estimate of mean between-group variance in math achievement test scores (i.e., sample mean of g^b_c) by small versus regular/regular-with-aide class types. Column 3 reports the difference in between-group variance across the two class types. Row 2 reports an estimate of 'expected' between-group variance (i.e., sample mean of g^w_c) by class type, with column 3 again reporting the difference. Both g^b_c and g^w_c are modified slightly to account for the fact that valid test scores are not observed for all students in every classroom. The ratio of the column 3 differences in rows 1 and 2 equals the Wald-IV estimate for γ^2, which is reported in row 3, column 3. Panel B repeats the exercise for reading achievement test scores. The first stage 'F-statistic' is simply the square of the t-statistic associated with the row 2, column 3 difference. Recall that y_{ci}, as used to compute g^b_c and g^w_c, is the residual associated with a preliminary regression of test scores on a vector of school dummy variables and the small class type dummy; this orthogonalization does not affect the computation of appropriate asymptotic standard errors.

Table 6: Tests for Social Interactions and 95 Percent Confidence Intervals for γ^2

	Math	**Reading**
Panel A : Wald Intervals/Tests for $\widehat{\gamma}^2$		
$W\left(\gamma_0^2\right) \quad \left(H_0\text{: } \gamma_0^2 = 1\right)$	4.07, $p = 0.0445$	2.51, $p = 0.1143$
$\left\{\gamma_0^2 \in \mathbb{R}^+ \middle\vert W\left(\gamma_0^2\right) < \chi_1^{2,0.95}\right\}$	(1.06, 5.12)	(0.31, 7.48)
Length	4.06	7.17
Right tail-to-left tail length ratio	1	1
Panel B : Wald Intervals/Tests for $\widetilde{\gamma} = \sqrt{\widehat{\gamma}^2}$		
$W\left(\gamma_0\right) \quad \left(H_0\text{: } \gamma_0 = 1\right)$	6.61, $p = 0.0106$	4.42, $p = 0.0364$
$\left\{\gamma_0 \in \mathbb{R}^+ \middle\vert W\left(\gamma_0\right) < \chi_1^{2,0.95}\right\}$	(1.18, 2.34)	(1.06, 2.89)
Length	1.16	1.83
Right tail-to-left tail length ratio	1	1
Panel C : ELR Intervals/Tests for $\widetilde{\gamma} = \sqrt{\widehat{\gamma}^2}$ & $\widehat{\gamma}^2$		
$LR\left(\gamma_0\right) \quad \left(H_0\text{: } \gamma_0 = 1\right)$	4.47, $p = 0.0344$	4.15, $p = 0.0417$
$\left\{\gamma_0 \in \mathbb{R}^+ \middle\vert LR\left(\gamma_0\right) < \chi_1^{2,0.95}\right\}$	(1.07, 2.31)	(1.05, 3.07)
Length	1.24	2.02
Right tail-to-left tail length ratio	0.81	1.18
$\left\{\gamma_0^2 \in \mathbb{R}^+ \middle\vert LR\left(\gamma_0^2\right) < \chi_1^{2,0.95}\right\}$	(1.15, 5.34)	(1.10, 9.42)
Length	4.18	8.32
Right tail-to-left tail length ratio	1.16	1.97

Notes: The Wald intervals reported in panel A are based on the normal approximation to the sampling distribution $\widehat{\gamma}^2$; those in panel B are based on the delta approximation for the sampling distribution of $\widetilde{\gamma} = \sqrt{\widehat{\gamma}^2}$. The empirical likelihood confidence intervals reported in panel C are based on the χ^2 approximation to the sampling distribution of the (profiled) empirical likelihood saddle-point criterion function (c.f., Newey and Smith 2004). This statistic, unlike the Wald one, is invariant to one-to-one parameter transformations.

Table 7: Peer Groups and Difference in Standardized Math and Reading Test Scores

	Math Score Change	Reading Score Change
Above average vs. below average student	1.1409	1.1573
	(0.0181)	(0.0269)
Classroom of above average vs. below average students	0.8643	1.1272
	(0.3376)	(0.5361)
Above average vs. below average Project STAR classroom	0.2023	0.2628
	(0.0790)	(0.1246)
Above average vs. below average teacher (upper bound estimate)	0.4046	0.3932
	(−)	(−)
Small vs. Regular/Regular-with-aide classroom	0.1631	0.1452
	(0.0466)	(0.0431)

Notes: Calculations for rows 1 to 4 are derived from GMM estimate of $\beta = \left(\varsigma\ \gamma^2\ \sigma^2\ \mu_{1/M}\right)'$ based on the moment function (23). Normality of the underlying α_c and ε_{ci} distributions is also assumed. The row 5 calculation is based on the regression results reported in Table 15. All standard errors computed using the delta method.

Table 8: Estimated Effects of Student Sorting on Achievement Inequality

Hypothetical Experiment	Math σ_y. Ratio	Reading σ_y. Ratio	Math $E[\Delta y_{ci}]$	Reading $E[\Delta y_{ci}]$
Modest to no sorting	1.0853	1.1243	0	0
	(0.0403)	(0.0761)	(−)	(−)
Medium to no sorting	1.2385	1.3386	0	0
	(0.1060)	(0.1916)	(−)	(−)
Perfect to no sorting	1.6672	1.9078	0	0
	(0.2625)	(0.4482)	(−)	(−)
Eliminate black/white gap in ε_{ci}	0.9495	0.9628	0.2674	0.2551
	(0.0151)	(0.0119)	(0.0609)	(0.0711)
Eliminate free lunch gap in ε_{ci}	0.9304	0.9177	0.3781	0.4541
	(0.0100)	(0.0124)	(0.0696)	(0.1117)

Notes: Columns 1 and 2 report estimates of the of standard deviation of math and reading achievement in the stated counterfactual relative to those actually observed in the Project STAR dataset (a random assignment benchmark). Columns 3 and 4 report estimates of the change in mean math and reading achievement associated with each counterfactual. Standard errors are reported in parentheses and computed by applying the delta method to the large sample variance-covariance matrix of the GMM estimates of $\beta^* = \left(\beta,\ \eta_b\ \eta_p\right)'$ based on the moment function (23). Modest, medium and perfect sorting correspond to $\zeta_{\varepsilon\varepsilon}$ equal to 0.1, 0.3, 1.0 respectively. Full details for each hypothetical experiment are provided in the main text.

Table 9: Estimates of γ^2 Based on Excess Variance Contrasts across Small and Medium versus Medium and Large Classrooms

	Small/Medium (WALD)	**Medium/Large (WALD)**	**Combined (GMM)**
Panel A : Math Achievement			
γ^2	4.030	−0.8482	3.1789
	(1.4282)	(3.5575)	(1.1277)
$F_{(df1,df2)} \quad 1^{st}-Stage$	$28.20_{(1,219)}$	$8.20_{(1,212)}$	$25.45_{(2,314)}$
$H_0 : \gamma^2_{s/m} = \gamma^2_{m/l}$	–	–	$p = 0.2488$
Panel B : Reading Achievement			
γ^2	4.5312	0.2144	3.2251
	(2.5477)	(2.7112)	(1.7330)
$F_{(df1,df2)} \quad 1^{st}-Stage$	$9.14_{(1,219)}$	$6.08_{(1,212)}$	$12.34_{(2,314)}$
$H_0 : \gamma^2_{s/m} = \gamma^2_{m/l}$	–	–	$p = 0.2828$
N	221	214	317

Notes: The small, medium, and large designations of classrooms are as described in the text. Estimation for columns 1 and 2 follows the procedure used in Table 5, using data from the appropriate subsample. Column 3 reports two-step GMM estimates on γ^2 based on the regression $g^b_c = \varsigma + \gamma^2 g^w_c$ with the three class size dummies serving as instruments. The null that $\gamma^2_{s/m} = \gamma^2_{m/l}$ is tested using the Sargan-Hansen test of overidentifying restrictions associated with the Column 3 estimates.

Table 10: Robustness to Upward Bias from Substitutibility between Teacher Quality and Class Size

	Math $\xi\left(\sigma^2_\alpha(0)\|\widehat{\gamma}^2, \widehat{\phi}^w_2, \gamma^2 = 1\right)$	**Reading** $\xi\left(\sigma^2_\alpha(0)\|\widehat{\gamma}^2, \widehat{\phi}^w_2, \gamma^2 = 1\right)$
$\sigma_\alpha(0) = 0.1$	2.40	2.50
$\sigma_\alpha(0) = 0.3$	1.24	1.26
$\widehat{\gamma}^2$	3.0891	3.8967
$\widehat{\phi}^w_2$	0.0228	0.0182

Notes: Rows 1 and 2 report the degree of substitutability between teacher quality and class size that would be required to produce the γ^2 estimates reported in Table 5 assuming two different values for standard deviation of teacher effectiveness in regular-sized classrooms, $\sigma_\alpha(0)$. The values of $\widehat{\phi}^w_2 = E\left[\widehat{g^w_c|q_c} = 1\right] - E\left[\widehat{g^w_c|q_c} = 0\right]$ and $\widehat{\gamma}^2$ used in the calibration are from rows 2 and 3, column 3, of Table 5.

Table 11: Estimates of γ^2 Based on Excess Variance Contrasts across High and Low Experience Heterogeneity Subsamples

	$Std\,(Exp_c) \geq 5$ **(WALD)**	$Std\,(Exp_c) < 5$ **(WALD)**	$Combined$ **(GMM)**
Panel A : Math Achievement			
Constant	0.0250	−0.0198	−0.0214
	(0.0701)	(0.0389)	(0.0357)
$Std\,(Exp_c) \geq 5$ years	–	–	0.0491
			(0.0214)
γ^2	3.0683	2.9586	3.0002
	(1.8388)	(1.1359)	(0.9922)
$F_{(df1,df2)}\quad 1^{st}-Stage$	$12.45_{(1,138)}$	$55.20_{(1,175)}$	$33.80_{(2,313)}$
$H_0 : \gamma^2_{HH} = \gamma^2_{LH}$	–	–	$p = 0.9593$
Panel B : Reading Achievement			
Constant	−0.0901	−0.0052	−0.0780
	(0.1342)	(0.0503)	(0.0760)
$Std\,(Exp_c) \geq 5$ years	–	–	0.0831
			(0.0405)
γ^2	6.5031	2.1251	3.7781
	(4.2795)	(1.1520)	(1.7280)
$F_{(df1,df2)}\quad 1^{st}-Stage$	$7.03_{(1,138)}$	$9.35_{(1,175)}$	$8.19_{(2,313)}$
$H_0 : \gamma^2_{HH} = \gamma^2_{LH}$	–	–	$p = 0.2966$
N	140	177	317

Notes: Columns 1 and 2 report Wald estimates of γ^2 based on subsamples exhibiting 'high' and 'low' degrees of heterogeneity in years of teacher experience. The construction of the two subsamples is described in the main text. Let HH_c= 1 if the c^{th} classroom is in the high heterogeneity subsample and zero otherwise (LH refers to low heterogeneity). Column 3 reports two-step GMM estimates on γ^2 based on the regression $g^b_c = \varsigma_{LH} + (\varsigma_{HH} - \varsigma_{LH}) \cdot HH_c + \gamma^2 g^w_c$ where g^w_c is instrumented with the small class type dummy and its interaction with HH_c. The null that $\gamma^2_{LH} = \gamma^2_{HH}$ is tested using the Sargan-Hansen test of overidentifying restrictions associated with the Column 3 estimates.

Table 12: Tests for Heterogenous Class Size Effects, Sorting, and Relative Bias Estimates

	Math	**Math**	**Reading**	**Reading**
Dependent Variable $M_c \cdot g_c^w$	NLS	NLS-RES	NLS	NLS-RES
σ_ω	0.8005	0.8189	0.8284	0.8508
	(0.0235)	(0.0142)	(0.0359)	(0.0252)
ρ_1	1.0798	1.0825	1.0165	1.0215
	(0.0373)	(0.0376)	(0.0517)	(0.0495)
ζ_L	−0.0692	–	−0.0835	–
	(0.0702)		(0.1148)	
$\frac{\widehat{\gamma}^2-\gamma^2}{\gamma^2}$ $\left(\text{Relative Bias of } \widehat{\gamma}^2\right)$	0.7109	–	0.1213	–
	(1.9105)		(1.6164)	
R^2	0.7654	0.7647	0.5956	0.5948
N	317	317	317	317

Notes: Nonlinear least squares estimates of σ_ω^2, ρ_1, and ζ_L based on equation (27) for math and reading achievement respectively. The fourth row reports estimates of the relative bias of an estimate of γ^2 based only on classrooms located in schools with more than three classrooms. Columns 2 and 4 report estimates with ζ_L constrained to zero. Reported standard errors calculated via a percentile bootstrap with 1,000 replications ($E\left[M_c^{-1}|q_c{=}\,1\right]$ and $E\left[M_c^{-1}|q_c{=}\,0\right]$ are set equal to their sample analogs and assumed to be non-stochastic).

Table 13: Implied Values of γ^2 for Different Signal-to-Noise Ratios

Classical Measurement Error	(1) $\kappa = 1.00$	(2) $\kappa = 0.90$	(3) $\kappa = 0.80$	(4) $\kappa = 0.60$	(5) $\kappa = 0.40$
MATH SCORES ($\widetilde{\gamma}^2_{EIV}$)	3.0891	3.3212	3.6114	4.4818	6.2228
	(1.0357)	(1.1508)	(1.2946)	(1.7262)	(2.5893)
READING SCORES ($\widetilde{\gamma}^2_{EIV}$)	3.8967	4.2186	4.6209	4.8278	8.2418
	(1.8294)	(2.0327)	(2.2868)	(3.0491)	(4.5735)

Notes: Rows 1 and 2 use equation (28) and the estimates of γ^2 reported in Table 5 to correct for classical measurement error in y_{ci} of varying intensities.

Table 14: Measurement-Error-Corrected Estimates of γ^2

	GMM
ς	-0.0405
	(0.0505)
γ^2	5.0609
	(2.1625)
σ^2	0.4819
	(0.0189)
$\sigma^2 + \sigma^2_{\upsilon\_\text{Math}}$	0.7363
	(0.0342)
$\sigma^2 + \sigma^2_{\upsilon\_\text{Reading}}$	0.7069
	(0.0204)
γ	2.2496
	(0.4806)
$\kappa_{\text{Math}} = \frac{\sigma^2}{\sigma^2+\sigma^2_{\upsilon 1}}$	0.6817
	(0.0160)
$\kappa_{\text{Reading}} = \frac{\sigma^2}{\sigma^2+\sigma^2_{\upsilon 2}}$	0.6545
	(0.0165)
N	317

Notes: GMM estimates of $\beta = \left(\varsigma,\ \gamma^2,\ \sigma^2,\ \sigma^2{+}\sigma^2_{\upsilon 1},\ \sigma^2{+}\sigma^2_{\upsilon 2}\right)$ based on the moment function (29). Standard errors for the social multiplier, γ, and the implied signal-to-noise ratios for the two tests, κ_1 and κ_2, are recovered via the delta method.

Table 15: OLS Conditional Mean Reduced Form Linear-in-Means Model of Social Interactions Estimates for Normalized Kindergarten SAT Math and Reading Scores

Reduced Form	(1) *OLS* Math	(2) *OLS* Reading
STUDENT-LEVEL		
π_{BLACK}	−0.3710	−0.2467
	$(0.0537)^{**}$	$(0.0546)^{**}$
π_{GIRL}	0.1311	0.1595
	$(0.0228)^{**}$	$(0.0249)^{**}$
$\pi_{FREELUNCH}$	−0.4243	−0.4611
	$(0.0283)^{**}$	$(0.0285)^{**}$
π_{DOB}	−0.2848	−0.1974
	$(0.0357)^{**}$	$(0.0357)^{**}$
CLASS MEANS		
$\pi_{\overline{BLACK}}$	0.1133	−0.5974
	(0.4950)	(0.4200)
$\pi_{\overline{GIRL}}$	0.3610	0.2788
	$(0.1835)^{*}$	$(0.1711)^{+}$
$\pi_{\overline{FREELUNCH}}$	−0.1060	0.0073
	(0.2030)	(0.1776)
$\pi_{\overline{DOB}}$	−0.2022	0.1625
	(0.2723)	(0.2432)
CLASS-LEVEL		
π_{SMALL}	0.1631	0.1452
	$(0.0466)^{**}$	$(0.0431)^{**}$
$\pi_{REGAIDE}$	−0.0182	−0.0392
	(0.0447)	(0.0394)
$\pi_{BLACKTEACHER}$	0.0344	0.0361
	(0.0791)	(0.0841)
$\pi_{MASTERS}$	−0.0290	−0.0073
	(0.0441)	(0.0409)
$\pi_{EXPERIENCE}$	0.0077	0.0095
	$(0.0042)^{+}$	$(0.0039)^{*}$
School fixed effects	*Yes*	*Yes*
$F_{(df1,df1)}$ for H_0: $\pi_{\overline{r}}=0$	$1.20_{(4,316)}$	$1.63_{(4,316)}$
	$p = 0.3117$	$p = 0.1670$
N	317	317
R^2	0.2805	0.2747
(n_y, n_w)	(5724, 6172)	(5646, 6172)

Notes: Rows '**', '*', and '+' denote that the reported coefficient is significantly different from zero at the 1, 5 and 10 percent level respectively. Reported standard errors are heteroscedastic robust with clustering at the classroom level. All regressions include school fixed effects; n_y denotes the total number of students with valid kindergarten test score data, while n_w denotes the total number of students in the $N = 317$ included classrooms regardless of test score status. Observations for all students are used to compute the observed peer composition variables. This avoids an error-in-variables problem that is described by Manski (1993) and Graham and Hahn (2004).

Table 16: Variable-by-Variable Tests for Excess Sensitivity in Normalized Kindergarten SAT Math and Reading Scores

Excess Sensitivity	(1) π_b	(2) π_w	(3) $\pi_b - \pi_w$
PANEL A : Math			
π_{BLACK}	−0.2577	−0.3710	0.1133
	(0.4861)	(0.0537)	(0.4950)
π_{GIRL}	0.4921	0.1311	0.3610
	(0.1842)	(0.0228)	$(0.1835)^{*}$
$\pi_{FREELUNCH}$	−0.5283	−0.4243	−0.1060
	(0.2008)	(0.0283)	(0.2030)
π_{DOB}	−0.4871	−0.2848	−0.2022
	(0.2693)	(0.0357)	(0.2723)
Omnibus Test Results	*Wald Statistic*	*df*	*p-value*
χ^2_K $(H_0 : \pi_{\overline{r}} = \pi_{\widetilde{r}})$	1.20	4	0.3117
χ^2_K $(H_0 : \pi_{\overline{r}} = 2 \cdot \pi_{\widetilde{r}})$	1.09	4	0.3608
PANEL B : Reading			
π_{BLACK}	−0.8441	−0.2467	−0.5974
	(0.4092)	(0.0546)	(0.4200)
π_{GIRL}	0.4384	0.1595	0.2788
	(0.1696)	(0.0249)	$(0.1711)^{+}$
$\pi_{FREELUNCH}$	−0.4538	−0.4611	0.0073
	(0.1773)	(0.0285)	(0.1776)
π_{DOB}	−0.0349	−0.1974	0.1625
	(0.2399)	(0.0357)	(0.2432)
Omnibus Test Results	*Wald Statistic*	*df*	*p-value*
χ^2_K $(H_0 : \pi_{\overline{r}} = \pi_{\widetilde{r}})$	1.63	4	0.1670
χ^2_K $(H_0 : \pi_{\overline{r}} = 2 \cdot \pi_{\widetilde{r}})$	2.13	4	0.0770

Notes: Columns 1 and 2 report the least squares coefficients on the between-and within-group transforms of the individual-level regressors entering the reduced form (see Table 15). Column 3 reports the difference in these two sets of coefficients variable-by-variable. '**','*', and "+" denote significance of these differences at the 1, 5 and 10 percent levels respectively. Reported standard errors are heteroscedastic robust with clustering at the classroom level. The 'Omnibus Test Results' panel reports Wald tests for the stated multi-coefficient restriction along with degrees-of-freedom and asymptotic p-values.

Table 17: Maximum Likelihood Estimates of the Linear-in-Means Model Based on Covariance Restrictions

	Math	**Reading**
Panel A : ML Estimates		
σ^2	0.7045	0.7227
	(0.0136)	(0.0140)
$\frac{\sigma_\alpha^2}{\sigma^2}\frac{1}{(1-\beta)^2}$	0.0051	−0.0043
	(0.0620)	(0.0542)
γ^2	3.0576	2.8132
	(1.2307)	(1.0816)
$LogL$	-7,293.7	-7,244.7
N	317	317
(n_s, n_o)	(5,724, 6,172)	(5,646, 6,172)
p-values for LR Test of H_0: $\gamma^2 = 1$	0.0668	0.0737

Notes: Likelihood is based on the assumption of joint normality of $(\underline{\varepsilon}_c', \alpha_c)' | w_c$. Standard errors are computed using the conditional Fisher Information. See the web appendix for more details on the likelihood and its maximization.

Table 18: Power of Excess Sensitivity and Variance Tests in Repeated Samples Calibrated to Mimic the Project STAR Dataset

	Math Power	*Inner inverse* γ^{II} $[\beta]$	*Outer inverse* γ^{OI} $[\beta]$	*Reading Power*	*Inner inverse* γ^{II} $[\beta]$	*Outer inverse* γ^{OI} $[\beta]$
Panel A : Power						
Excess Sensitivity	0.8469	1.52 [0.34]	1.88 [0.47]	0.7309	1.53 [0.35]	1.91 [0.48]
Excess Variance	0.9937	1.38 [0.28]	1.63 [0.39]	0.9808	1.37 [0.27]	1.62 [0.38]
Relative Odds	28.51			18.81		
Panel B : Calibration						
σ^2	0.7045			0.7227		
$\rho_\alpha = \sigma_\alpha^2/\sigma^2$	0.0017			0.0000		
$\eta'\Sigma_{rr}\eta$	0.0568			0.0545		
$\sigma_\epsilon^2 = \sigma^2 - \eta'\Sigma_{rr}\eta$	0.6477			0.6682		
$\gamma = 1/(1-\beta)$	1.7486			1.6773		
β	0.43			0.40		

Notes: Power approximations computed using the inverse CDF of a non-central χ^2 random variable with the non-centrality parameters given the text. The inner and outer inverse power functions are computed as described by Andrews (1989).