TESTING RANDOM ASSIGNMENT TO PEER GROUPS

KOEN JOCHMANS*
UNIVERSITY OF CAMBRIDGE

May 1, 2020

Abstract

Identification of peer effects is complicated by the fact that the individuals under study may self-select their peers. Random assignment to peer groups has proven useful to sidestep such a concern. In the absence of a formal randomization mechanism it needs to be argued that assignment is 'as good as' random. This paper introduces a simple yet powerful test to do so. We provide theoretical results for this test and explain why it dominates existing alternatives. Asymptotic power calculations and an analysis of the assignment mechanism of players to playing partners in tournaments of the Professional Golfer's Association is used to illustrate these claims. Our approach can equally be used to test for the presence of peer effects. To illustrate this we test for the presence of peer effects in the classroom using kindergarten data collected within Project STAR. We find no evidence of peer effects once we control for classroom fixed effects and a set of student characteristics.

Keywords: asymptotic power, bias, peer effects, random assignment.

JEL classification: C12, C21.

*Address: University of Cambridge, Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DD, United Kingdom. E-mail: `kj345@cam.ac.uk`.
Financial support from the European Research Council through grant n° 715787 (MiMo) is gratefully acknowledged.
The Stata command `rassign` implements the test developed here and can be installed from within Stata by typing `ssc install rassign` in the command window. I am most grateful to Vincenzo Verardi for help in the development of this command.

Introduction

A fundamental issue when trying to infer peer effects is the concern that the individuals under study, at least partially, self-select their reference group. Exploiting the random assignment of individuals to peer groups has proven to be a fruitful way forward. Sacerdote (2001) and Zimmerman (2003) estimate peer effects in college achievement by making use of the (conditional) random assignment of students to roommates. Katz, Kling and Liebman (2001) and Duflo and Saez (2003) are other early examples that use such exogenous variation in other settings.

In many studies on peer effects there is no formal randomization mechanism. In others the randomization is done at a higher level than under the experimental ideal. Examples of the former situation are in the work of Bandiera, Barankay and Rasul (2009) and Mas and Moretti (2009), both of which concern workers being assigned to teams or shifts. An example of the latter is Project STAR, where students appear to have been randomly assigned only to classes of a certain size, not to classrooms themselves; see Sojourner (2013) for a detailed discussion on this. In such settings more work is needed to convincingly argue that the assignment of peers is 'as good as random'.

Sacerdote (2001) pioneered a regression-based approach to test for random assignment. Guryan, Kroft and Notowidigdo (2009) pointed out that this test favors alternatives where there is negative assortative matching between peers, and suggested a modification.[1] Their proposal has been used frequently—Carrell, Fullerton and West (2009), Sojourner (2013), and Lu and Anderson (2015) are examples—but it has not been subject to theoretical investigation. The limited simulation evidence available suggests that it is size correct but has low power (Stevenson, 2015). Thus, the test would have difficulty in detecting

[1]The intuition given in Guryan, Kroft and Notowidigdo (2009) and repeated elsewhere in the literature (Caeyers and Fafchamps, 2020) is that individuals cannot be their own peers. While this argument explains why the test favors negative alternatives it does not explain the cause of the size distortion. In fact, minor modifications to the proof of (1.1) below show that size distortion would also be present when individuals can be their own peers. Furthermore, in such a case the test will tend to favor alternatives where assortative matching is positive. In all cases, the cause of the (asymptotic) size distortion is the presence of fixed effects.

violations of the null of random assignment.

In this paper we propose an alternative adjustment to the test of Sacerdote (2001), and study its properties under the null and under various local alternatives. The approach is based on a bias calculation and is straightforward to implement (a Stata implementation is also available). It allows both peer groups and urns from which peers are drawn to be of the same or of different sizes, accommodates designs in which peer groups need not be mutually exclusive, and is robust to heteroskedasticity of arbitrary form. Because assignment is usually random only conditional on allocation to urns, our test, like Sacerdote's (2001), controls for fixed effects at the urn level. A straightforward modification to the test that allows to control for additional covariates is also presented.

The derivations underlying our test also allow to establish formal results for the test of Guryan, Kroft and Notowidigdo (2009). First, we confirm that the test is indeed size correct. Moreover, their proposal corresponds to an alternative way of performing the bias correction that is inherent in our procedure, when either an urn-level homoskedasticity assumption is satisfied or peer groups are mutually exclusive. This alternative approach is only implementable when there is variation in urn size, however. Second, we provide an asymptotic representation that helps to explain the low power that has been observed for the test of Guryan, Kroft and Notowidigdo (2009). We illustrate the power loss through theoretical power calculations and show that the test can have trivial power against a wide range of alternatives. In all cases considered our test is uniformly more powerful than theirs, and considerably so.

The test developed here can equally be applied to test for the presence of peer effects in the linear-in-means model without modification. This is a useful observation because the test does not require the usual conditions for identification in such settings under the alternative. Furthermore, identification is much easier to establish once such effects can be ruled out.

We present two empirical applications of our test that illustrate its usefulness. The first is a re-analysis of the data on professional golf tournaments of Guryan, Kroft and Notowidigdo (2009). Here, players that enter a tournament are randomly assigned to playing partners, conditional on belonging to the same player category. Like theirs, our test supports that this is indeed the case. However, unconditional on player categories, player assignment is non-random. While our test convincingly detects this violation, the test of Guryan, Kroft and Notowidigdo (2009) continues to strongly support the null of random assignment. This type-II error is a direct consequence of the test having low power.

To illustrate an alternative use of our test, our second empirical illustration tests for the presence of peer effects in student performance. We use the data on SAT mathematics scores of kindergarten students in 317 Tennessee classrooms collected within Project STAR. The data from Project STAR have been analysed extensively for a variety of purposes. Graham (2008) and Rose (2017) use the same data as do we to estimate models of peer effects. While identification can be achieved through information contained in second moments of test scores there is a concern that in the Project STAR data it is weak (see Rose 2017, p. S55 for a discussion). Our approach is different. Rather than fitting an unrestricted model we test for the presence of peer effects directly. If such effects can be ruled out, the problem of identification simplifies considerably. In our data, there is evidence of such effects conditional only on classroom fixed effects. However, once we additionally control for a set of characteristics this significance vanishes. Hence, we do not find evidence of spillover effects here.

The paper is organized as follows. Section 1 sets up the problem, derives our test statistic, and presents its statistical properties. Section 2 connects to the alternative tests proposed elsewhere and, notably, provides a theoretical comparison to the proposal of Guryan, Kroft and Notowidigdo (2009). Section 3 contains two extensions. First, to allow for arbitrary heteroskedasticity; these calculations also verify that our original test is fully robust to heteroskedasticity when peer groups are mutually exclusive. Second, It also shows how to modify the approach to accommodate additional control variables. Section 4 presents our two empirical illustrations. A short conclusion end the paper. All proofs are collected in the Appendix.

1 Testing random assignment

Consider a setting where we observe stratified data on r independent urns containing, respectively, $n_1, \dots n_r$ individuals. Within each urn individuals are assigned to peer groups. The assignment of peers in urn g is recorded in the $n_g \times n_g$ matrix

$$(\boldsymbol{A}_g)_{i,j} := \begin{cases} 1 & \text{if } i \text{ and } j \text{ are peers} \\ 0 & \text{if they are not} \end{cases};$$

as individuals cannot be their own peer matrix $\boldsymbol{A}_g$ has only zeros on its main diagonal.[2] The number of peers of individual i is $m_g(i) := \sum_{j=1}^{n_g} (\boldsymbol{A}_g)_{i,j}$. We assume that each individual has at least one peer but do not otherwise restrict peer groups; they may be of different sizes and are allowed to overlap. The goal is to test whether individuals are randomly assigned to their respective peer groups.

Let $x_{g,i}$ be an observable characteristic of individual i in urn g. Sacerdote (2001) noted that, under random assignment, $x_{g,i}$ will be uncorrelated with $x_{g,j}$ for all $j \in [i]$, where $[i] := \{j : (\boldsymbol{A}_g)_{i,j} = 1\}$ is the set of i's peers. Letting $\bar{x}_{g,[i]} := m_g(i)^{-1} \sum_{j=1}^{n_g} (\boldsymbol{A}_g)_{i,j}\, x_{g,j}$, the average value of the characteristic among i's peers, he then proceeded by testing whether the slope coefficient in a within-group regression of $x_{g,i}$ on $\bar{x}_{g,[i]}$ is statistically different from zero. The within-group estimator controls for fixed effects at the urn level. This is important as, even if assignment is randomized within urns, individuals might be assigned to an urn based on other attributes. In the data of Sacerdote (2001), for example, students are randomly assigned to rooms conditionally on gender and their answers to a set of survey questions. If peer assignment within urns is presumed to only be random conditional on a set of additional covariates $\boldsymbol{w}_{g,i}$, say, they can equally be controlled for by including them as additional regressors.

[2] Everything to follow can be modified to deal with situations where the adjacency matrices $\boldsymbol{A}_1, \dots, \boldsymbol{A}_r$ are asymmetric (as in directed networks), have non-binary entries (covering weighted networks), and have a non-zero main diagonal (allowing individuals to be their own peer). To maintain focus we do not pursue the most general case here.

1.1 Bias calculation

As observed by Guryan, Kroft and Notowidigdo (2009), the test just described will typically not be size correct. To see the problem, and a path forward, we start by a bias calculation. For now we ignore any additional covariates $\boldsymbol{w}_{g,i}$ and thus consider a fixed-effect regression of $x_{g,i}$ on $\bar{x}_{g,[i]}$. The within-group estimator, $\hat{\rho}$, is defined as the solution to the normal equation

$$\sum_{g=1}^{r}\sum_{i=1}^{n_g} \bar{x}_{g,[i]} \left(\tilde{x}_{g,i} - \hat{\rho}\,\tilde{\bar{x}}_{g,[i]}\right) = 0,$$

where $\tilde{x}_{g,i}$ and $\tilde{\bar{x}}_{g,[i]}$ are deviations of, respectively, $x_{g,i}$ and $\bar{x}_{g,[i]}$ from their within-urn mean. A calculation given in the Appendix shows that the normal equation is biased. Moreover,

$$\mathbb{E}_0\left(\sum_{g=1}^{r}\sum_{i=1}^{n_g} \bar{x}_{g,[i]}\,\tilde{x}_{g,i}\right) = -\sum_{g=1}^{r}\sigma_g^2, \tag{1.1}$$

where the subscript on the expectations operator indicates that the expectation is taken under the null of random assignment, and we have assumed that $\mathbb{E}_0(\tilde{x}_{g,i}^2) =: \sigma_g^2$ does not vary across individuals. This urn-level homoskedasticity assumption can be dispensed with and we do so below. Furthermore, it will turn out that, when peer groups are mutually exclusive, the test derived under this homoskedasticity assumption is, in fact, robust to heteroskedasticity.

Equation (1.1) implies that the within-group estimator is inconsistent under asymptotics where the number of urns grows large but their size is held fixed. In the Appendix we show that (under the null)

$$\operatorname{plim}_{r\to\infty}\hat{\rho} = -\frac{\lim_{r\to\infty}\frac{1}{r}\sum_{g=1}^{r}\sigma_g^2}{\lim_{r\to\infty}\frac{1}{r}\sum_{g=1}^{r}\sigma_g^2\;\mathbb{E}_0\left(\sum_{i=1}^{n_g}\frac{1}{m_g(i)} - \frac{1}{n_g}\sum_{i=1}^{n_g}\sum_{j=1}^{n_g}\frac{m_g(i\cap j)}{m_g(i)\,m_g(j)}\right)}, \tag{1.2}$$

where $m_g(i\cap j) := \sum_{k=1}^{n_g}(\boldsymbol{A}_g)_{i,k}\,(\boldsymbol{A}_g)_{k,j}$ is the number of peers that individuals i and j have in common. The probability limit is always negative. All else equal its magnitude is decreasing in urn sizes and increasing in the degree of overlap between peer groups. When peer groups do not overlap it is also increasing in the size of the peer groups. Furthermore, in the special case where all urns are of size n and are partitioned into peer groups of a common size m,

$$\operatorname{plim}_{r\to\infty} \hat{\rho} = -\frac{m}{n-1},$$

which no longer depends on the urn variances. This expression co-incides with the one reported in Proposition 1 of Caeyers and Fafchamps (2020).

The implication of the inconsistency is that the regression-based test will be biased toward negative alternatives and that its size will tend to one as the number of urns grows large.

1.2 A corrected test

The bias calculated in (1.1) is surprisingly simple and suggests a natural adjustment to the test statistic of Sacerdote (2001). Observe that an unbiased estimator of σ_g^2 (under the null) is

$$\frac{1}{n_g - 1} \sum_{i=1}^{n_g} x_{g,i}\, \tilde{x}_{g,i}.$$

Therefore, the re-centered covariance

$$q_r^{\mathrm{HO}} := \sum_{g=1}^{r} \sum_{i=1}^{n_g} \bar{x}_{g,[i]}\, \tilde{x}_{g,i} + \sum_{g=1}^{r} \frac{1}{n_g - 1} \sum_{i=1}^{n_g} x_{g,i}\, \tilde{x}_{g,i} = \sum_{g=1}^{r} \sum_{i=1}^{n_g} \tilde{x}_{g,i} \left(\bar{x}_{g,[i]} + \frac{x_{g,i}}{n_g - 1} \right)$$

will be exactly unbiased under random assignment. An estimator of the standard deviation of q_r^{HO} is a conventional standard error that clusters observations at the urn level. It equals

$$s_r^{\mathrm{HO}} := \sqrt{\sum_{g=1}^{r} \left(\sum_{i=1}^{n_g} \tilde{x}_{g,i} \left(\bar{x}_{g,[i]} + \frac{x_{g,i}}{n_g - 1} \right) \right)^2}.$$

Hence, an adjusted test statistic is $t_r^{\mathrm{HO}} := q_r^{\mathrm{HO}} / s_r^{\mathrm{HO}}$. Note that the entire construction of this statistic is based on calculations under the null. As such it is in the spirit of a Lagrange-multiplier test.[3]

Theorem 1 states the asymptotic behavior of the statistic t_r^{HO} under the null and under alternatives where $\mathbb{E}(s_r^{\mathrm{HO}}) = b_r$ for a sequence of constants $b_r = O(\sqrt{r})$. Illustrations of Pitman drifts of this type are given below.

[3] Note that t_r^{HO} can equally be viewed as a convential t-statistic—obtained through a bias-corrected within-group regression—that uses a standard error that is constructed under the null.

Theorem 1. *Let* $\mathbb{P}(n_g > 2) = 1$. *If* $\max_{g,i} \mathbb{E}(x_{g,i}^8) = O(1)$ *and* $\max_{g,i}(\mathbb{E}(x_{g,i}^2))^{-1} = O(1)$, *then*

$$t_r^{\text{HO}} - \frac{b_r}{s_r^{\text{HO}}} \xrightarrow{d} N(0,1),$$

as $r \to \infty$.

It is easy to verify that urns of size two would not contribute to the test statistic and so can be dropped. Hence the need for the first condition in the theorem. The second condition contains standard moment requirements.

An implication of the theorem is that, for any $\alpha \in (0,1)$,

$$\lim_{r\to\infty} \mathbb{P}_0 \left(t_r^{\text{HO}} > z_{1-\alpha}\right) = \alpha,$$

where z_α is the α-quantile of the standard-normal distribution. One-sided and two-sided tests then follow in the usual manner. The theorem also implies that the test is consistent against any alternative for which b_r does not grow slower than $\sqrt{r}$. We turn to such deviations next.

The bias adjustment in q_r^{HO} is smaller for urns of larger size. This may suggest that in settings where peers are drawn from large urns, ignoring the bias issue in the test of Sacerdote (2001) is inconsequential (Guryan, Kroft and Notowidigdo, 2009). Such reasoning ignores the fact that the standard deviation of q_r^{HO}, too, is decreasing in urn sizes. The conclusion, in line with results in the panel data literature (e.g., Hahn and Kuersteiner 2002), is that the bias will only be ignorable for testing purposes when the size of the urns is substantially larger than the number of urns. We note, though, that in such a case the usual cluster-robust variance estimator should not be used. Alternative variance estimators are provided in Stock and Watson (2008).

1.3 Power calculations

We consider three types of local alternatives, where $x_{g,i}$ is correlated across peers. In the terminology of Manski (1993) these are (i) endogenous effects, (ii) contextual effects, and (iii) correlated effects. We begin by providing a closed-form expression for the variance of

q_r^{HO} under the null. We then calculate b_r under the alternatives (i)–(iii). Taken together, these results then yield the non-centrality parameter in the limit distribution of t_r^{HO}. This is then used to assess power.

Throughout this subsection we focus attention on settings where peer groups do not overlap, which makes the final expressions more easily interpretable. We also enforce that $\mathbb{E}_0(x_{g,i}^4) = 3\sigma_g^4$, which yields a slightly shorter variance formula but is in no way essential to our findings. The underlying derivations in the Appendix do not make use of these restrictions.

Variance expression. Under these conditions the variance of q_r^{HO} under the null is equal to

$$v_r^{\mathrm{HO}} := \mathbb{E}_0(q_r^{\mathrm{HO}}\, q_r^{\mathrm{HO}}) = 2 \sum_{g=1}^{r} \sigma_g^4 \; \mathbb{E}_0 \left(\sum_{i=1}^{n_g} \frac{1}{m_g(i)} - \frac{n_g}{n_g - 1} \right). \tag{1.3}$$

We observe that v_r^{HO} is increasing in the size of the urns and decreasing in the size of the peer groups.

Endogenous effects. In our first set of alternatives correlation among peers arises through

$$x_{g,i} = \rho\, \bar{x}_{g,[i]} + \varepsilon_{g,i}, \qquad \varepsilon_{g,i} \sim \text{independent}\,(\alpha_g, \sigma_g^2),$$

where $-1 < \rho < 1$ and the $\varepsilon_{g,i}$ are independent of the matrix $\boldsymbol{A}_g$. A drifting sequence of this model towards the null is obtained by setting $\rho = \varrho/\sqrt{r}$ for fixed values of ϱ. Such local alternatives imply that

$$b_r = 2 \frac{\varrho}{\sqrt{r}} \sum_{g=1}^{r} \sigma_g^2 \; \mathbb{E} \left(\sum_{i=1}^{n_g} \frac{1}{m_g(i)} - \frac{n_g}{n_g - 1} \right). \tag{1.4}$$

Note that this term depends on the design in the same way as does v_r^{HO} and so the same comparative statistics apply. Taken together, by an application of Theorem 1, t_r^{HO} will converge in distribution to a normal random variable with mean $\mu := \lim_{r\to\infty} b_r/\sqrt{v_r^{\mathrm{HO}}}$ and variance one. The larger μ (in magnitude) the smaller the probability of a type-II error. The non-centrality parameter μ is even simpler when errors are homoskedastic and the adjacency matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r$ are drawn from a common distribution as, in that case,

$$\mu = \varrho \sqrt{2\,\mathbb{E}\left(\sum_{i=1}^{n_g} \frac{1}{m_g(i)} - \frac{n_g}{n_g - 1}\right)},$$

showing that power is monotone increasing in the (expected) size of the urns and decreasing in the size of the peer groups. When variances are urn specific the expression for μ is to be multiplied by

$$\lim_{r\to\infty} \frac{1}{\sqrt{r}} \frac{\sum_{g=1}^r \sigma_g^2}{\sqrt{\sum_{g=1}^r \sigma_g^4}} \leq 1,$$

where the bound follows from the Cauchy-Schwarz inequality. Hence, urn-specific variances are always power reducing. Nonetheless, note that $\mu > 0$, and so our test will detect endogenous-effect violations with probability approaching one for all possible configurations of urn sizes and peer groups.

Contextual effects. In our second class of alternatives correlation in peer characteristics comes from (latent) exogenous effects. Moreover,

$$x_{g,i} = \varepsilon_{g,i} + \frac{\theta}{m_g(i)} \sum_{j=1}^{n_g} (\boldsymbol{A}_g)_{i,j}\, \varepsilon_{g,j}, \qquad \varepsilon_{g,i} \sim \text{independent}\,(\alpha_g, \sigma_g^2)$$

where θ is a finite constant and, again, the $\varepsilon_{g,i}$ are independent of the matrix $\boldsymbol{A}_g$. For drifting sequences of the form $\theta = \vartheta/\sqrt{r}$,

$$b_r = 2\frac{\vartheta}{\sqrt{r}} \sum_{g=1}^{r} \sigma_g^2\, \mathbb{E}\left(\sum_{i=1}^{n_g} \frac{1}{m_g(i)} - \frac{n_g}{n_g - 1}\right), \tag{1.5}$$

which is the identical to the bias under an endogenous-effect alternative where $\varrho = \vartheta$. Consequently, endogenous and exogenous effects are locally asymptotically equivalent. This finding is not surprising in light of the similar results on autoregressive and moving-average alternatives in classical testing problems in the time series literature (see, for example, Godfrey 1981).

Correlated effects. In our third class of alternatives peers are subject to a common additive shock drawn from a distribution with variance σ_η^2, independent of everything else. Thus (conditional on an urn fixed effect) the variance of $x_{g,i}$ is equal to $\sigma_\eta^2 + \sigma_g^2$ while the covariance between characteristics of peers is σ_η^2. In this case, the relevant drifting sequence has $\sigma_\eta^2 = \varsigma^2/\sqrt{r}$ and we find that the bias in q_r^{HO} equals

$$b_r = \frac{\varsigma^2}{\sqrt{r}} \sum_{g=1}^{r} \mathbb{E}\left((n_g - 1) - \frac{1}{n_g} \sum_{i=1}^{n_g} \frac{m_g(i)}{n_g - 1} \right). \tag{1.6}$$

Because $\sum_{i=1}^{n_g} m_g(i) \leq n_g(n_g - 1)$, with equality if and only if all individuals in urn g are each others peers we again have that $b_r > 0$ and so our test will be consistent against all correlated-effect alternatives. When $\sigma_g^2 = \sigma^2$ and the matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r$ are drawn from a common distribution, the non-centrality parameter in the limit distribution of our test statistic is

$$\mu = \frac{\varsigma^2}{\sigma^2} \frac{\mathbb{E}\left((n_g - 1) - \frac{1}{n_g} \sum_{i=1}^{n_g} \frac{m_g(i)}{n_g - 1} \right)}{\sqrt{2\,\mathbb{E}\left(\sum_{i=1}^{n_g} \frac{1}{m_g(i)} - \frac{n_g}{n_g - 1} \right)}}.$$

Power is again increasing in $n_1, \ldots, n_r$. The impact of the size of the peers groups on power is less clear cut, however. On the one hand, larger peer groups reduce the variance and increase μ. On the other hand, they also reduce the bias in q_r^{HO}, resulting in a loss of power.

2 Connections to the literature

When there is variation in urn size Guryan, Kroft and Notowidigdo (2009) proposed to augment the within-group regression of Sacerdote (2001) by including the leave-one-out average

$$\frac{1}{n_g - 1} \sum_{j \neq i} x_{g,j} = \frac{n_g}{n_g - 1} \left(\frac{1}{n_g} \sum_{j=1}^{n_g} x_{g,j} - \frac{x_{g,i}}{n_g} \right) = \frac{n_g}{n_g - 1} \left(\overline{x}_g - \frac{x_{g,i}}{n_g} \right)$$

as an additional regressor. The within-group transformation sweeps out all terms that do not vary within urns, and so the approach is equivalent to a within-group regression of $x_{g,i}$ on $\bar{x}_{g,[i]}$ and $x_{g,i}/(n_g - 1)$. This highlights why variation in urn size is required. When

n_g does not vary across urns this regression will yield a perfect fit that satisfies the null whether or not peer assignment is random. Guryan, Kroft and Notowidigdo (2009) offer an intuition of why their strategy yields size control and provide supporting simulations. However, a theoretical analysis of the test is, to our knowledge, not available.

Calculations summarized in the Appendix reveal that the approach of Guryan, Kroft and Notowidigdo (2009) tests whether

$$\sum_{g=1}^{r}\sum_{i=1}^{n_g} \tilde{x}_{g,i}\left(\bar{x}_{g,[i]} + \frac{x_{g,i}}{n_g - 1}\right)\left(1 - \frac{\delta}{n_g - 1}\right) + o_p(\sqrt{r}), \tag{2.7}$$

is statistically different from zero. Here,

$$\delta := \frac{\lim_{r\to\infty} \frac{1}{r}\sum_{g=1}^{r} \sigma_g^2}{\lim_{r\to\infty} \frac{1}{r}\sum_{g=1}^{r} \sigma_g^2 \, \mathbb{E}_0\left(\frac{1}{n_g-1}\right)},$$

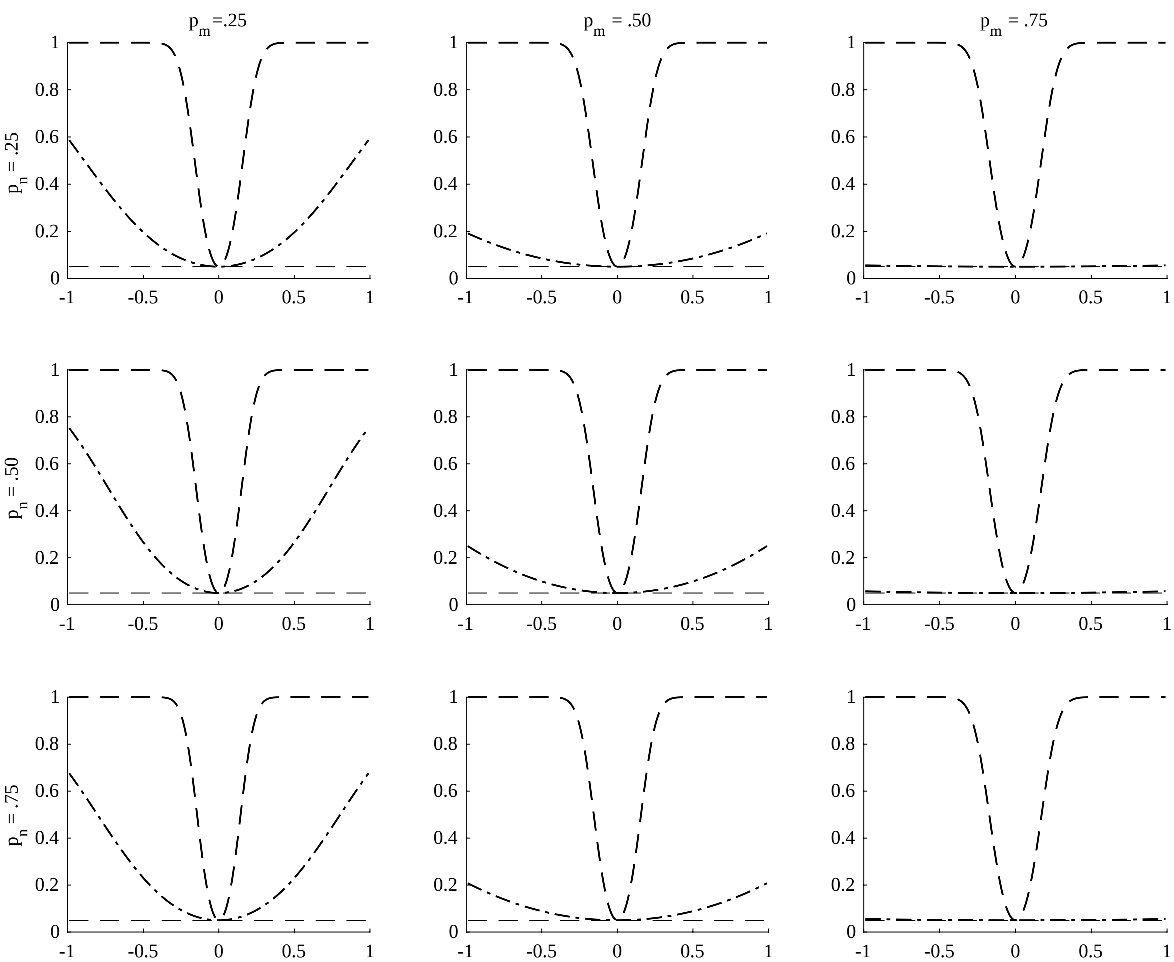
is the probability limit of the slope coefficient of a within-group regression of $x_{g,i}$ on $x_{g,i}/(n_g - 1)$, under the null. The summand in the leading term in (2.7) is equal to the summand in q_r^{HO}, up to a scale factor that varies at the urn level. This factor is bounded and so, by virtue of Theorem 1, we conclude that the test will indeed exhibit correct size in large samples.

The limited simulation evidence available suggests that the test of Guryan, Kroft and Notowidigdo (2009) may suffer from low power; see Stevenson (2015) and also the extended version of her analysis in the Appendix. Because the approach requires variation in urn sizes one may expect the test to be particularly underpowered when such variation is limited (Stevenson 2015, Caeyers and Fafchamps 2020). While this is true, low power can also arise from a different source. Equation (2.7) is again useful here. Consider a design where urns are of size $\bar{n}_1$ with probability $(1 - p_n)$ and of size $\bar{n}_2$ with probability p_n, where $\bar{n}_1 < \bar{n}_2$. Then the non-centrality parameter of the test statistic can be shown to equal

$$\mu^* := \sqrt{p_n(1-p_n)}\frac{b(\bar{n}_2) - b(\bar{n}_1)}{\sqrt{v(\bar{n}_1)\, p_n + v(\bar{n}_2)\,(1-p_n)}}, \tag{2.8}$$

where $b(n)$ and $v(n)$ are the bias and variance of $\sum_{i=1}^{n_g} \tilde{x}_{g,i}\,(\bar{x}_{g,[i]} + x_{g,i}/(n_g - 1))$ conditional on $n_g = n$. This equation confirms that $\mu^* \to 0$ as $p_n(1-p_n) \to 0$ and formalizes the notion that the test will tend to have low power when variation in urn sizes is small. The formula also shows that the test will have trivial asymptotic power when $b(\bar{n}_1) - b(\bar{n}_2) = 0$, i.e., in designs where the bias contributions coming from the different urn sizes cancel each other out.

Figure 1: Power analysis for endogenous- and exogenous-effect alternatives



Power against endogenous/exogenous effect alternatives for our test (dashed line) and for the test of Guryan, Kroft and Notowidigdo (2009) (dashed-dotted line) in a design with two possible urns sizes (4 and 6) and two possible peer-group sizes (2 and 3). $p_n := \mathbb{P}(n_g = 6)$ and $p_m := \mathbb{P}(m_g(i) = 2|n_g = 6)$. A horizontal dashed-dotted line indicates the size of the test. Plots are based on theoretical calculations and are for 25 urns.

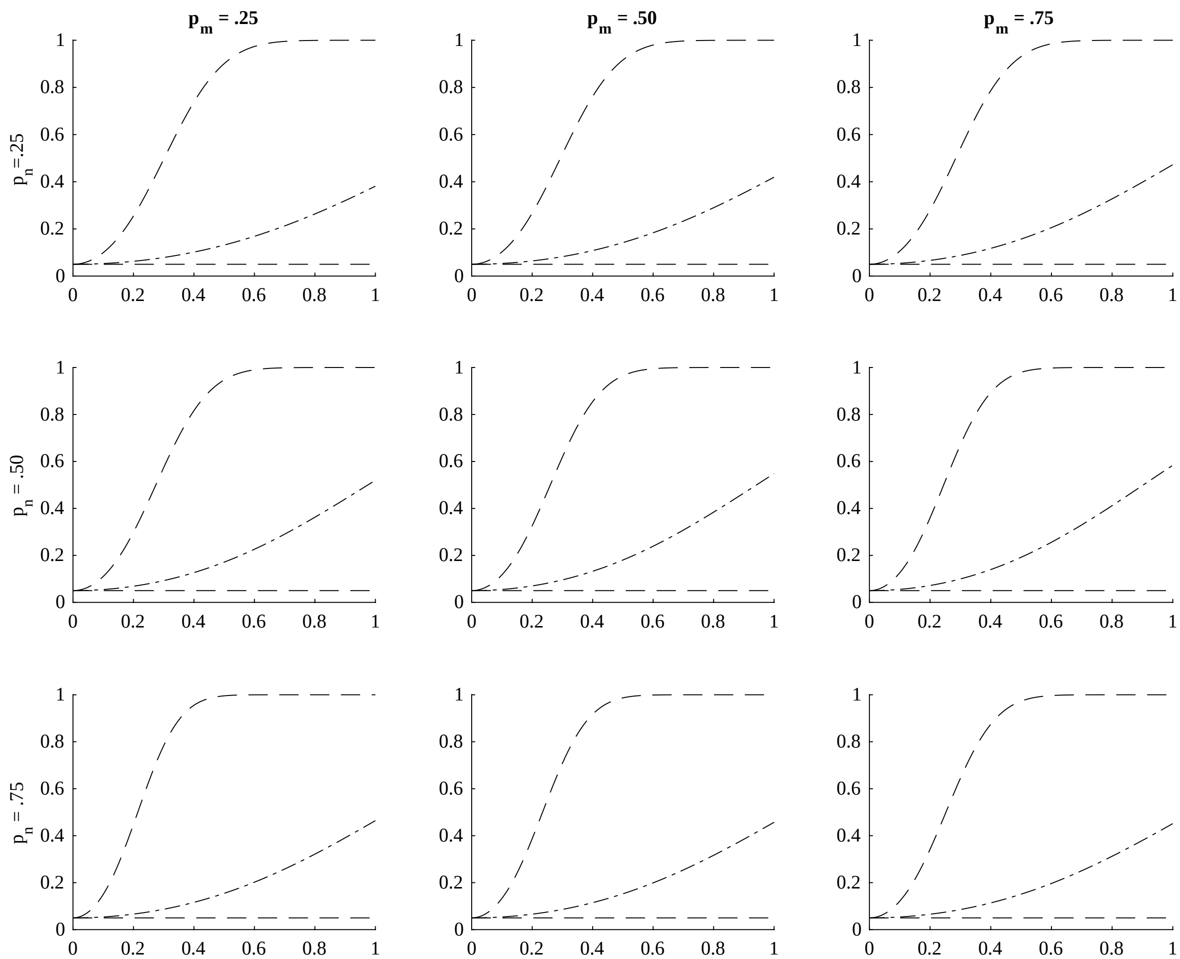
We confirm these findings in Figures 1 and 2 for designs where each of 25 urns contains six individuals with probability p_n and four individuals with probability $1 - p_n$. Within earns of size four, each individual is assigned one peer at random while in the larger urn peer groups are of size three with probability p_m and of size two with probability $1 - p_m$. Figure 1 plots (theoretical) power against endogenous (or, equivalently, contextual) effect alternatives, with ρ (or, equivalently, θ) on the horizontal axis. Figure 2 displays power against correlated-effect alternatives, with σ_η^2/σ^2 on the horizontal axis. The plots in each figure are arranged so that p_n increases when going down rows and p_m increases when moving through columns. Dashed curves refer to our test. Dashed-dotted curves represent the test of Guryan, Kroft and Notowidigdo (2009). Both tests are two-sided at the 5% level; a dashed horizontal line marks the size.

Figure 1 shows high power for our test across all designs. The test of Guryan, Kroft and Notowidigdo (2009) is uniformly less powerful, and substantially so. There is a reduction in its power when p_n moves away from .50 (i.e., across rows). For the values considered here, this effect is small relative to the impact of changing p_m, with power initially going down considerably when p_m moves from .25 to .50, and afterwards essentially flattening out completely when $p_m = .75$. This is a reflection of the numerator in μ^* getting close to zero; the bias in urns of size four cancels out with the bias in urns of size six. As μ^* is multiplicative in ρ these changes are uniform on $(-1, 1)$.

Figure 2 shows our test also has high power against correlated-effect alternatives. The power gain in the test of Guryan, Kroft and Notowidigdo (2009) as σ_η^2/σ^2 moves further away from zero (the null) trails behind considerably. However, in contrast to the pattern in Figure 1, we do not observe trivial power in any of the configurations. The reason for this is that, here, for none of the combinations of p_n and p_m the numerator of μ^* is close to zero. A close look will allow to verify that, here, power increases with p_m. This is in line with our formulas.

Guryan, Kroft and Notowidigdo (2009) also describe an alternative permutation test (see, e.g., Lehmann and Romano 2006, Chapter 15, for a general treatment of such tests) that is based on the sampling distribution of the within-group estimator obtained from randomly re-assigning individuals to peer groups within each urn. Randomization tests

Figure 2: Power analysis for correlated-effect alternatives



Power against correlated-effect alternatives for our test (dashed line) and for the test of Guryan, Kroft and Notowidigdo (2009) (dashed-dotted line) in a design with two possible urns sizes (4 and 6) and two possible peer-group sizes (2 and 3). $p_n := \mathbb{P}(n_g = 6)$ and $p_m := \mathbb{P}(m_g(i) = 2|n_g = 6)$. A horizontal dashed-dotted line indicates the size of the test. Plots are based on theoretical calculations and are for 25 urns.

have many attractive properties but require that individuals are exchangeable under the null to be size correct. This is a substantial strenghtening of the requirements underlying Theorem 1. A relevant data feature that would violate exchangeability is when $x_{g,i}$ is heteroskedastic (in i), for example.

Stevenson (2015) suggested an alternative approach based on data splitting. Although its theoretical properties have not been established, the subsampling scheme proposed circumvents bias under the null, at least when peer groups are mutually exclusive, and so should lead to size correct inference in this case (under regularity conditions). Like the permutation test, the scheme is also computationally much more demanding than our bias-adjustment proposal.

3 Extensions

3.1 Heteroskedasticity

So far we have worked under an assumption of urn-level homoskedasticity. We now drop this restriction and allow that $\sigma^2_{g,i} := \mathbb{E}_0(x^2_{g,i})$ varies both between and within urns in an arbitrary way.

First, calculations analogous to those that gave rise to (1.1) show that, now,

$$\mathbb{E}_0\left(\sum_{g=1}^{r}\sum_{i=1}^{n_g} \bar{x}_{g,[i]}\,\tilde{x}_{g,i}\right) = -\sum_{g=1}^{r} \mathbb{E}_0\left(\frac{1}{n_g}\sum_{i=1}^{n_g}\frac{1}{m_g(i)}\sum_{j=1}^{n_g}(\boldsymbol{A}_g)_{i,j}\,\sigma^2_{g,j}\right). \tag{3.9}$$

Hence, the contribution of each urn to the bias equals (minus) the expected within-urn mean of peer-group averaged variances.

Appealing to a result of Hartley, Rao and Kiefer (1969), we show in the Appendix that an unbiased estimator of the bias in (3.9) is

$$-\sum_{g=1}^{r}\sum_{i=1}^{n_g}\omega_{g,i}\,x_{g,i}\,\tilde{x}_{g,i}, \qquad \omega_{g,i} := \frac{1}{n_g-2}\left(\sum_{i'\in[i]}\frac{1}{m_g(i')} - \frac{1}{n_g-1}\right),$$

which is again well-defined for all urns of size $n_g > 2$. Hence, a modification of q_r^{HO} that is robust to heteroskedasticity of arbitrary form is given by

$$q_r^{\mathrm{HC}} := \sum_{g=1}^{r}\sum_{i=1}^{n_g}\tilde{x}_{g,i}\,\left(\bar{x}_{g,[i]} + \omega_{g,i}\,x_{g,i}\right), \tag{3.10}$$

which satisfies $\mathbb{E}_0(q_r^{\mathrm{HC}}) = 0$. It differs from q_r^{HO} only in that the weight $(n_g-1)^{-1}$ is replaced by $\omega_{g,i}$, which varies at the individual level. Construction of $\omega_{g,i}$ is nonetheless immediate from $\boldsymbol{A}_g$.

Observe that, in the important special case where peer groups do not overlap we have $m_g(i') = m_g(i)$ for all $i' \in [i]$, and so

$$\omega_{g,i} = \frac{1}{n_g - 1}.$$

This is the weight we used to construct our test statistic in the homoskedastic case. It thus follows that t_r^{HO} is robust to heteroskedasticity in this case.

The standard deviation of q_r^{HC} can be estimated by

$$s_r^{\mathrm{HC}} := \sqrt{\sum_{g=1}^{r} \left(\sum_{i=1}^{n_g} \tilde{x}_{g,i} \, \left(\bar{x}_{g,[i]} + \omega_{g,i} \, x_{g,i} \right) \right)^2}.$$

A modified version of our test statistic that remains size correct under heteroskedasticity of arbitrary form also when peer groups overlap is $t_r^{\mathrm{HC}} := q_r^{\mathrm{HC}} / s_r^{\mathrm{HC}}$. This statistic is asymptotically normal under the same conditions as before. In the following theorem, $b_r := \mathbb{E}(q_r^{\mathrm{HC}}) = O(\sqrt{r})$.

Theorem 2. *Let $\mathbb{P}(n_g > 2) = 1$. If $\max_{g,i} \mathbb{E}(x_{g,i}^8) = O(1)$ and $\max_{g,i} (\mathbb{E}(x_{g,i}^2))^{-1} = O(1)$, then*

$$t_r^{\mathrm{HC}} - \frac{b_r}{s_r^{\mathrm{HC}}} \xrightarrow{d} N(0,1),$$

as $r \to \infty$.

3.2 Controlling for covariates

There may be situations where, in addition to urn fixed effects, it is desirable to control for other variables that vary at the individual level, $\boldsymbol{w}_{g,i}$. This would be needed when randomization is assumed to take place within urns only conditional on these variables. A intuitive regression-based solution would be to first partial-out $\boldsymbol{w}_{g,i}$ from $x_{g,i}$ and then proceed in constructing our test statistic as before. We next show that, under regularity conditions, this approach is justified.

Let $\dot{x}_{g,i}$ denote the residual from an ordinary least-squares regression of $x_{g,i}$ on urn dummies and the vector of covariates $\boldsymbol{w}_{g,i}$. Then the modified test statistic takes the form

$$\hat{t}_r^{\mathrm{HO}} := \frac{\hat{q}_r^{\mathrm{HO}}}{\hat{s}_r^{\mathrm{HO}}}$$

for

$$\hat{q}_r^{\mathrm{HO}} := \sum_{g=1}^{r} \sum_{i=1}^{n_g} \dot{x}_{g,i} \left(\bar{x}_{g,[i]} + \frac{x_{g,i}}{n_g - 1} \right), \qquad \hat{s}_r^{\mathrm{HO}} := \sqrt{\sum_{g=1}^{r} \left(\sum_{i=1}^{n_g} \dot{x}_{g,i} \left(\bar{x}_{g,[i]} + \frac{x_{g,i}}{n_g - 1} \right) \right)^2}.$$

The statistic t_r^{HC} can be modified in the same way.

To state conditions under which Theorem 1 generalizes to partialling-out covariates we need

$$\check{x}_{g,i} := x_{g,i} - \boldsymbol{w}'_{g,i} \left(\sum_{g=1}^{r} \sum_{i'=1}^{n_g} \mathbb{E}(\boldsymbol{w}_{g,i'} \boldsymbol{w}'_{g,i'}) \right)^{-1} \left(\sum_{g=1}^{r} \sum_{i'=1}^{n_g} \mathbb{E}(\boldsymbol{w}_{g,i'} x_{g,i'}) \right).$$

This is the deviation of $x_{g,i}$ from its population linear projection on $\boldsymbol{w}_{g,i}$ (and no fixed effects).

The following theorem provides the result. Here, $\|\cdot\|$ refers to the Euclidean norm and b_r is once more suitably re-defined to be the bias in $\hat{q}_r^{\mathrm{HO}}$ under Pitman drifts towards the null hypothesis.

Theorem 3. *Let* $\mathbb{P}(n_g > 2) = 1$. *If* $\max_{g,i} \mathbb{E}(\check{x}_{g,i}^8) = O(1)$ *and* $\max_{g,i} (\mathbb{E}(\check{x}_{g,i}^2))^{-1} = O(1)$, *then*

$$\hat{t}_r^{\mathrm{HO}} - \frac{b_r}{\hat{s}_r^{\mathrm{HO}}} \xrightarrow{d} N(0,1),$$

as $r \to \infty$, *provided that* $\mathbb{E}(\check{x}_{g,i} | \boldsymbol{w}_{g,1}, \ldots, \boldsymbol{w}_{g,n_g}) = \alpha_g$ *for urn-specific constants* $\alpha_1, \ldots, \alpha_r$, *that* $\max_{g,i} \mathbb{E}(\|\boldsymbol{w}_{g,i}\|^4) = O(1)$ *and that the matrix* $\lim_{r\to} r^{-1} \sum_{g=1}^{r} \mathbb{E}(\tilde{\boldsymbol{w}}_{g,i} \tilde{\boldsymbol{w}}'_{g,i})$ *has maximal rank.*

The conditions in this result are intuitive. First, the moment conditions on $x_{g,i}$ in Theorem 1 are replaced by corresponding conditions on $\check{x}_{g,i}$. Next, the mean-independence assumption is a requirement of strict exogeneity on $\boldsymbol{w}_{g,i}$. Finally, the conditions on the covariates are needed to ensure that the residuals from the auxiliary least-squares regression converge to their population counterparts.

4 Illustrations

4.1 Randomization in professional golf tournaments

Guryan, Kroft and Notowidigdo (2009) used the random assignment of golf players to playing partners in professional golf tournaments to estimate peer effects. Their data span the 2002, 2005, and 2006 seasons of the Professional Golfer's Association (PGA) and cover 81 tournaments. We refer to Guryan, Kroft and Notowidigdo (2009) for a detailed description of the data. Here we only note the facts that are of direct relevance to our analysis. Players in the PGA are, at any point in time, assigned to one of four categories (cat 1, cat 1a, cat 2, and cat 3). At the start of each tournament, within these four categories, playing partners are assigned to groups of three golfers. These (mutually exclusive) peer groups play together for the first two rounds of the tournament. The analysis is limited to the first round. Conditional on the set of players who enter a tournament, the assignment is random within categories. Unconditional on this fully interacted set of fixed effects, assignment to groups is not random (Guryan, Kroft and Notowidigdo, 2009, p. 40). Random assignment is tested by looking at the (corrected) within-group correlation between a measure of a golfer's ability and the average ability of his peers in the reference group.

The chief measure of ability used to do this is an estimate of the number of strokes more than 72 (i.e., above par) that a golfer typically takes in a round, on an average course, that is used for PGA tournaments. The more negative this number the better the player. Table 1 contains descriptive statistics for this variable, stratified by the four player categories. It shows that, broadly, average ability is higher in lower numbered categories, and that there remains substantial variation in this measure even conditional on category. To get a sense of urn sizes in these data the table also provides descriptive statistics of the number of players by tournament (tourn) and by tournament-by-category. The latter are based on a total of 8,791 observations in stead of the total of 8,801 observations as 10 observations concern urns of a size less than three; recall that such urns do not contain any information for our purposes. We also included the same descriptive statistics for the weights $(n_g - 1)^{-1}$.

Table 1: The PGA data

	n obs	mean	std	min	max
ability ($x_{g,i}$)					
cat 1	3,205	-3.138	0.769	-5.159	1.440
cat 1a	3,436	-2.808	0.740	-4.326	6.732
cat 2	1,503	-2.857	0.894	-4.776	3.275
cat 3	657	-1.662	1.470	-4.776	6.315
peer ability ($\bar{x}_{g,[i]}$)					
cat 1	3,205	-3.132	0.599	-5.081	0.672
cat 1a	3,436	-2.811	0.591	-4.530	3.275
cat 2	1,503	-2.850	0.744	-4.776	3.275
cat 3	657	-1.690	1.270	-4.776	6.315
urn size (n_g)					
tourn	8,801	111.942	18.414	62	144
tourn by cat	8,791	39.292	16.869	3	83
weight ($(n_g - 1)^{-1}$)					
tourn	8,801	0.009	0.002	0.007	0.016
tourn by cat	8,791	0.037	0.040	0.012	0.500

The test statistics for the default (i.e., uncorrected) regression-based test, our corrected version, and the test where leave-me-out urn means are controlled for are collected in Table 2. The numbers in square brackets below are corresponding (two-sided) p-values. When fully stratifying the data by tournament and category we observe that the default test rejects the null of random assignment and would suggest there to be negative assortative matching between players. The other two tests have large p-values, finding little evidence to contradict the null. Recall that the assignment of players to groups is not random when not controlling for categories. We would hope that both tests capture this violation from the null. Our corrected test does this; its p-value drops from .394 to .000. The test of Guryan, Kroft and Notowidigdo (2009) on the other hand continues to suggest that golfers are randomly assigned; its p-value actually increases from .227 to .329. This type-II error is in line with our theoretical results on this test.

Table 2: Results for the PGA data

stratification	default	corrected	control
tourn	7.524	4.288	-0.976
	[0.000]	[0.000]	[0.329]
tourn by cat	-3.957	-0.852	-1.209
	[0.000]	[0.394]	[0.227]

We conclude this illustration by highlighting a caveat to the analysis of these data. Most, if not all, professional golf players participate to multiple tournaments per year and are also active for multiple years. Consequently, many players will appear in multiple urns, albeit with a different value for their ability measure, as this is updated over time. This, of course, induces dependence across urns which is in violation with our working assumption that urns are independent.

4.2 Peer effects in the classroom

We use data collected as part of Project STAR to test for the presence of peer effects among kindergarten students. These data are well known and have been used extensively.

The data set we is borrowed from Graham (2008). It covers 317 kindergarten classrooms in the state of Tennessee. A summary of the data is given in Table 3. We have SAT scores for mathematics taken at the end of the year (math), and dummies for gender (girl), ethnicity (black), and eligibility for free school meals (lunch). The SAT scores are standardized to have mean zero and unit variance. On entering kindergarten students were randomly assigned to one of three class types. There has been debate on whether students were also randomly assigned to classes; see Graham (2008), Chetty et al. (2011), and Sojourner (2013). The concensus seems to be that violations appear small, especially at the kindergarten level.

Table 3: The Project STAR data

	n obs	mean	std	min	max
math score	5,724	0.000	1.000	-4.129	2.943
peer math	5,724	0.000	0.581	-1.616	2.123
girl	5,724	0.486	0.500	0	1
black	5,724	0.327	0.469	0	1
lunch	5,724	0.480	0.500	0	1
class size	317	18.057	3.965	9	27

Table 4 provides the value of our test statistic for the null of no peer effects for three different specifications. For completeness it also contains its uncorrected counterpart. The baseline specification controls only for classroom fixed effects. Here, the null is strongly rejected. The second specification additionally controls for the observed characteristics of the students. In this case, we no longer find evidence of peer effects. The same is true if we augment the control variables by the average characteristics of the peers. Thus, we do not find evidence of any type of spillover effects once background characteristics are controlled for.

An alternative to our direct test would be to estimate the linear-in-means model and explicitely test for the presence of endogenous and exogenous effects. Because classrooms do not overlap this cannot be done through the popular instrumental-variable approach

Table 4: Results for the Project STAR data

	default	corrected	default	corrected	default	corrected
test statistic	-56.710	4.215	-55.710	-1.295	-56.085	1.211
p-value (two sided)	0.000	0.000	0.000	0.195	0.000	0.226
class fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
controls	No	No	Yes	Yes	Yes	Yes
controls of peers	No	No	No	No	Yes	Yes

of Bramoullé, Djebbari and Fortin (2009) and De Giorgi, Pellizzari and Redaelli (2010) that hinges on partial overlap between peer groups. Exploiting the variation in class size, Rose (2017) (extending the work of Graham 2008) used the identifying power in second moments of test scores to back out estimates of endogenous and exogenous effects in our data. Both point estimates (in his Specification 3 in Table 2) came out as insignificant at all conventional significance levels. This result is in line with the conclusion obtained here. It is noted in Rose (2017, p. S55) that identification of the full model in the Project STAR setting may be weak. This aids in rationalizing the large standard errors he obtains. It also highlights the usefulness of a test for the presence of peer effects that circumvents the need for a consistent estimator of the unrestricted model.

Conclusion

Random assignment of individuals to peer groups has proven to be a powerful tool for credible identification of spillover effects. In non-experimental designs it needs to be argued that assignment is 'as good as' random. This paper has presented a simple test to do so. Its properties were derived and a comparison to alternative test available the literature has been made.

Peer groups may be of different sizes and need not be mutually exclusive. Variation in the size of the urns from which peers are drawn is allowed but is not necessary. Individuals within urns are not assumed to be exchangeable under the null. Theoretical analysis verifies that the test is consistent against endogenous effects, contextual effects, and correlated effects. We also provide theoretical results that illustrate substantial power improvements over the test of Guryan, Kroft and Notowidigdo (2009). These calculations also clarify why this test will often have low power.

In a first empirical illustration we verify random assignment in the professional golfer data of Guryan, Kroft and Notowidigdo (2009). Within tournaments, participating players are randomly assigned to playing partners from the same category. Like theirs, our test confirms this. Assignment is not random when not controlling for categories. While our test successfully detects this violation, theirs does not.

In a second empirical example we use our approach to test for the presence of peer effects in educational achievement. Using data on SAT scores in kindergarten collected as part of Project STAR we do not find evidence of peer effects once student characteristics and classroom fixed effects are accounted for.

References

Bandiera, O., I. Barankay, and I. Rasul (2009). Social connections and incentives in the workplace: Evidence from personnel data. *Econometrica 77*, 1047–1094.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009). Identification of peer effects through social networks. *Journal of Econometrics 150*, 41–55.

Caeyers, B. and M. Fafchamps (2020). Exclusion bias in the estimation of peer effects. NBER Working Paper No. 22565.

Carrell, S. E., R. L. Fullerton, and J. E. West (2009). Does your cohort matter? Measuring peer effects in college achievement. *Journal of Labor Economics 27*, 439–464.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics 126*, 1593–1660.

De Giorgi, G., M. Pellizzari, and S. Redaelli (2010). Identification of social interactions through partially overlapping peer groups. *American Economic Journal: Applied Economics 2*, 241–275.

Duflo, E. and E. Saez (2003). The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *Quarterly Journal of Economics 118*, 815–842.

Godfrey, L. G. (1981). On the invariance of the Lagrange multiplier test with respect to certain changes in the alternative hypothesis. *Econometrica 49*, 1443–1455.

Graham, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica 76*, 643–660.

Guryan, J., D. Kroft, and N. J. Notowidigdo (2009). Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics 44*, 289–302.

Hahn, J. and G. Kuersteiner (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large. *Econometrica 70*, 1639–1657.

Hartley, H. O., J. N. K. Rao, and G. Kiefer (1969). Variance estimation with one unit per stratum. *Journal of the American Statistical Association 64*, 841–851.

Katz, L., J. Kling, and J. Liebman (2001). Moving to opportunity in Boston: Early results of a randomized mobility study. *Quarterly Journal of Economics 116*, 607–654.

Lehmann, E. L. and J. P. Romano (2006). *Testing Statistical Hypotheses*. Springer.

Lu, F. and M. Anderson (2015). Peer effects in microenvironments: The benefits of homogeneous classroom groups. *Journal of Labor Economics 33*, 91–122.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies 60*, 531–542.

Mas, A. and E. Moretti (2009). Peers at work. *American Economic Review 99*, 112–145.

Rose, C. (2017). Identification of peer effects through social networks using variance restrictions. *Econometrics Journal 20*, S47–S60.

Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics 116*, 681–704.

Sojourner, A. (2013). Identification of peer effects with missing peer data: Evidence from Project STAR. *Economic Journal 123*, 574–605.

Stevenson, M. (2015). Tests of random assignment to peers in the face of mechanical negative correlation: An evaluation of four techniques. Mimeo.

Stock, J. H. and M. W. Watson (2008). Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica 76*, 155–174.

Zimmerman, D. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and Statistics 85*, 9–23.