Peer Effects in University Housing: Evidence from Fuzzy Central Assignment*

Chao Fu† Jesse Gregory‡

Clint Harris§ Victoria Prowse¶

This version: February 10, 2026
First version: October 28, 2022

Abstract

We estimate effects of roommates, neighbors, and dorms on academic outcomes at a large public four-year university. To address selection, we instrument for realized room assignments with simulated offers generated by a room assignment mechanism, while controlling for expected offers implied by the same mechanism. Our candidate assignment mechanisms fail to perfectly replicate the assignment offers made by the central planner, so we select the mechanism that best replicates assignments via a data-driven model selection procedure. We find that inhabiting a fully gender-integrated dorm increases 4-year graduation by 10 percentage points for men, with no significant effects on women, while finding consistent (statistically insignificant) results for exposure to female neighbors within dorms. At the roommate level, we find that STEM roommates have significantly smaller negative effects on four-year graduation for other STEM students than they do for non-STEM students. Our findings suggest that universities with 70% gender-integrated housing (such as the university we study) could increase their male 4-year graduation rates by 3 percentage points by implementing full gender-integration in all dorms, with no offsetting negative effects on women.

Keywords: Peer Effects, College Majors, Central Assignment, Model Selection, Simulated Instruments

JEL Classification: C52, D47, I20

*Institutional Review Board approvals were obtained from The University of Wisconsin (2021-0422) and Purdue University (IRB-2021-1024).

†Department of Economics, University of Wisconsin-Madison; cfu@ssc.wisc.edu.

‡Department of Economics, University of Wisconsin-Madison; jmgregory@ssc.wisc.edu.

§Corresponding Author: Wisconsin School of Business, University of Wisconsin-Madison; clint.harris@wisc.edu.

¶Department of Economics, Purdue University; vprowse@purdue.edu.

1 Introduction

A large number of young people attend universities, many of whom live in university housing. Their assignments to dorms and peers are of policy interest because complementarities between individuals and their dorms or peers allows for the possibility of alternative living arrangements that may improve outcomes on average, or even for all groups. Because many aspects of these living arrangements are determined by university administrators according to partially random procedures, a large literature has developed to consider which types of peers positively affect other peers, with a particular focus on nonlinear effects which enable aggregate gains from counterfactual peer assignments. In this paper, we estimate effects of roommate, neighbor, and dorm characteristics on a range of academic outcomes in the setting of a public four-year university by leveraging intent-to-treat simulated peer assignments generated by a data-driven room assignment mechanism.

A large literature investigates effects of peers in university housing on academic and vocational outcomes. However, this literature overwhelmingly leverages data from private universities (Sacerdote, 2001; Stinebrickner and Stinebrickner, 2006; Marmaros and Sacerdote, 2006; Sacerdote, 2011) and military academies (Carrell et al., 2013, 2019; Jones and Kofoed, 2020) where the randomization of room assignments is well-understood, but which might not be representative of peer effects in more common settings. We provide evidence on effects of roommates, neighbors, and dorms in student housing at a large public four year institution in the United States, merging administrative data on pre-existing student characteristics, room assignments, and outcomes with internal-use data from University Housing on dorm and roommates preferences. Specifically, we follow 24,265 incoming freshmen at the University of Wisconsin-Madison from 2016-2019 who submit housing preferences the summer before their first year on campus.

It is possible that the limited attention to large public universities in the literature is due to room assignment complexities that invalidate commonly-applied empirical strategies that condition on observed room assignment strata indicators as in the research described above. For instance, Foster (2006) discusses significant complications regarding nonrandom roommate assignments at the University of Maryland, and focuses her intention instead on estimating effects of peers who inhabit the same wings of dorms, rather than rooms. We find that room assignment procedures in our public four year institutional setting are substantially more complex and less consistently documented than those described in the literature.[1] We overcome this challenge by applying and extending recent methodological innovations by Abdulkadiroğlu et al. (2017), Borusyak and Hull (2020), and others that instrument for treatment assignments with treatment offers, while controlling for expected values of offers. Similar methods have been used to estimate effects, for instance, of school value added (Angrist et al., 2017), charter schools (Abdulkadiroğlu et al., 2017), and travel distance to school (Angrist et al., 2022a). We are the first (to our knowledge) to apply these methods to estimation of peer effects in a university setting.

Unlike recent applications in the school choice literature of these methods, we are not able to replicate the observed room assignments with a parsimonious room assignment mechanism, eliminating our ability to control for expected values of treatment offers. Our key methodological insight is that the results of Abdulkadiroğlu et al. (2017) and Borusyak and Hull (2020) permit treatment effect estimation of simulated treatments when controlling for expected simulated treatments *regardless of whether simulated treatment offers correspond to realized treatment offers*. It

[1] For instance, in interviews with housing staff, we were informed that university housing staff exercises judgment while doing room assignments, intentionally attempting to improve roommate match quality according to unspecified rules. One example given was that they avoid assigning students from the same high school as roommates if the high school is small. There is neither a requirement that staff always perform this correction nor a standard metric for determining whether a high school is small.

follows that the dorm or peer assignments implied by any arbitrary room assignment mechanism are valid (but perhaps irrelevant) instruments for realized treatment assignments when controlling for students' expected dorm or peer assignments implied by the same mechanism. In other words, we construct "intended intention-to-treat instruments", where we as researchers intend for our assignment mechanism to produce treatment assignments that match those intended by the central assigner. Realized assignments may differ from our simulated assignments either because of lack of take-up from the central planner when they make assignment offers that conflict with those that we simulate, or because of lack of take-up from students when they refuse assignment offers made by the central planner.

While any room assignment mechanism is likely to produce conditionally independent instruments, there is no requirement that it will produce relevant instruments that reliably predict realized dorm or peer assignments. Formally, this constitutes a setting with infinitely many potential instruments and a finite number of observations. To construct relevant instruments, we simulate room assignments according to a small number of candidate mechanisms that satisfy parsimony constraints, and we test the accuracy of each mechanism against the assignment offers (not realized assignments). To avoid complicating inference for treatment effects via pretesting, we select the room assignment mechanism that best rationalizes the randomly-generated (and therefore independent of relevant peer traits) lottery room assignment lottery tiebreakers of students' peers implied by assignment offers, rather than attempting to match dorm or peer characteristic variables of realized peers directly. Model selection methods such as those we use are potentially useful to school (etc.) choice applications with researcher ignorance regarding true assignment mechanisms as well as in the more general framework of Borusyak and Hull (2020) where there is uncertainty regarding the economic model generating the data.

We estimate effects of roommate ACT scores, roommate STEM status, neighbor gender, and dorm-level gender integration. Our results on dorms are particularly well-powered, and are made possible by our school-choice style identification style that explicitly leverages a substantial amount of between-dorm identifying variation. We find that assignment to coeducational dorms (those with no partitions between genders) increases four-year graduation rates by over 10 percentage points for men (90% CI) while having no significant effect on women. Given that just under 70% of students in our sample live in coed dorms, linear extrapolationn suggests that our university could potentially increase four-year graduation for male students by 3 percentage points, with no negative offsets to women, by converting all dorms to be fully coeducational. This is consistent with previous findings that women and girls in educational settings positively effect college graduation rates (Hill, 2017), high school graduation rates and test scores (Lavy and Schlosser, 2011), and primary school reading and math scores (Gottfried and Harven, 2015).

In addition to effects at the dorm level, we estimate effects of neighbor gender by leveraging more granular variation in exposure to opposite-gender peers made possible by the existence of coeducational dorms. To do this, we manually geocode dormrooms in three dimensions using university blueprints. This allows us to construct a matrix of (Manhattan) distances in inches between each room and every other room on the same floor. We define neighbors for individuals in each room as those individuals that inhabit the nearest three rooms, allowing us to investigate effects of peers at a more granular level than past work, such as Foster (2006) who found null results of nearby peers when considering coarser wing-level variation in peers. We that exposure to female neighbors increases the probability of graduating with a STEM major by 19 percentage points, with suggestive (but insignificant) evidence that this effect is larger for men. We also find statistically insignificant evidence that exposure to female peers positively effects freshmen retention and four-year graduation for men with no effects on women — which we view as complementary to our similar estimates of effects of coeducational dorms.

We also consider effects of roommates. We estimate effects of peer STEM interest (based on intended majors), peer verbal ACT scores, peer math ACT scores, and distances between individuals values of these traits and those of their roommates. We find that assignment to a STEM roommate has a 7 percentage point more positive effect for STEM individuals relative to non-STEM individuals on four-year graduation rates, with opposite effects of similar magnitude on students graduating with a STEM major. We also find that a standard deviation (8.8 point) increase in ACT verbal score distance narrowly falls short (p-value 0.13) of statistical significance in negatively affecting freshmen retention by 7 percentage points. These findings broadly suggest potential benefits of clustering students together based on academic abilities and interests, rather than exposing them to academically diverse peers.

The paper proceeds as follows: Section 2 provides background on student housing at the University of Wisconsin; Section 3 describes the empirical strategy; Section 4 describes the data; Section 5 presents the results; and Section 6 concludes.

2 Background

Most US universities maintain on-campus housing for undergraduate students. This typically takes the form of several dorms (residence halls), each with dorm rooms spread over several floors. Dorms may be gender-integrated (coed) or gender-segregated and vary in their amenities, such as location and dining facilities. Dorm rooms are typically dual occupancy, but single triple- or even quad-occupancy dorm rooms also exist. The allocation of students to dorm rooms is generally managed by an office for student housing.

We study peer effects in university housing at the University of Wisconsin-Madison (henceforth 'the University'), a large, four-year public research university with an undergraduate population of 33,500 students in Fall 2021. We focus on undergraduate students who lived in the University's dorms between 2016 and 2019. During this time period, the University had 22 dorms for undergraduate housing, of which zero were single-sex, four had single-sex floors, four had single-sex wings or room clusters within floors, and fourteen had no partitions between rooms housing individuals of different genders (with the potential for randomly assigned opposite gender neighbors). Seven of the dorms that had mixed gender floors also had gender inclusive rooms, in which students of any gender could live with a predetermined roommate of any gender (students are never randomly assigned an opposite sex roommate). Thirteen of the dorms included wings or other groups of rooms that are reserved for students admitted to Learning Communities (e.g., Women in Science and Engineering, WISE). Nine dorms had single person rooms, twenty-two had two-person rooms, twelve had three-person rooms, eleven had four-person rooms, and three had six-person rooms, with the majority of students residing in two-person rooms. Between 2016 and 2019, over 90% of first-year undergraduate students lived in dorms, along with many higher-year students.

Division of University Housing at UW-Madison (known as 'Housing') is responsible for allocating incoming first-year and returning students to dorm rooms. Incoming students are advised to apply for housing before a cutoff date in the summer before the start of the fall semester, while returning students face an earlier cutoff date. The assignment mechanism is a proprietary procedure that is not fully documented. Through extensive interviews with Housing staff, we learned that the following procedure is used to allocate rooms to students who applied for housing before their cutoff date.

1. The University's athletics department assigns student-athletes to rooms.

2. The following groups of students choose their rooms in order: a) students admitted to Learn-

ing Communities (restricted to rooms in their Learning Community); b) returning students; c) incoming transfer students and first-year students above age twenty.

3. Housing assigns genders to remaining rooms to ensure that dorms have space proportional to incoming student gender shares (thereafter, students are only assigned to rooms that match their gender).

4. Housing uses a Random Serial Dictator (RSD) assignment rule to place unassigned students into dorms while rooms within dorms are allocated at the discretion of Housing staff:

   (a) Students complete a housing questionnaire that elicits their housing preferences. In particular, students are asked to list the University's dorms in order of preference (students may rank as many dorms as they wish). Students may also state their preferred room occupancy (single, double, triple or quad) and may designate a preferred roommate with whom they wish to share a room; if the preferred roommate submits the same housing preferences, the students become 'predetermined roommates.'

   (b) Each student is given a randomly generated 'tiebreaker number.' The tiebreaker number for students with predetermined roommates is replaced by the lowest tiebreaker number of the students in the roommate group. Students are then sorted in ascending order of the tiebreaker number.

   (c) Housing staff proceed through the sorted list from top to bottom. Each student (and any predetermined roommates) is assigned to their most preferred dorm among those dorms with remaining space and then assigned to a room within the assigned dorm. If all dorms on the student's preference list are full, Housing staff use discretion to place the student (and any predetermined roommates) in an alternative available dorm room. Housing does not have official guidelines for its staff regarding room placements within dorms but claims to respect room occupancy preferences where possible.[2]

5. Housing staff use their judgment to adjust dorm and room assignments made in Step 4 above, e.g., they may make adjustments to avoid students with particular preferences being clustered together.

Students who apply for housing after the cutoff date are placed into dorm rooms at the discretion of the Housing staff. After students are notified of their initial housing assignments, they may request revisions at any time.

Students assigned in Steps 1-3 of the housing allocation procedure will not be randomly assigned to dorms, but they may have randomly assigned peers (roommates or dorm neighbors), as the peers assigned to them may be assigned according to the RSD assignment rule describe in Step 4 above. Similarly, students with predetermined roommates will not have randomly assigned roommates but may still have randomly assigned neighbors and be randomly assigned to a dorm.

3 Empirical Strategy

For residential treatments indexed by k in $k = 1, 2, ..., K$ and individuals indexed by i, we are interested in the effect of a treatment D_{ik} (e.g. a binary indicator for living in a highly gender

[2]Empirically, we find that students clustered together in the random tiebreaker order are also clustered geographically within dorms.

integrated dorm) on an outcome Y_i (e.g. graduating from college). We model this effect as the coefficient β_k in the equation

$$Y_i = \beta_0 + \beta_k D_{ik} + \epsilon_i, \tag{1}$$

where β_0 represents a constant and ϵ_i contains unobserved determinants of the outcome. This model of the true effect of D_{ik} on Y_i can be estimated via (naive) OLS, which will produce an $\hat{\beta}_k^N$ that is consistent for β_k if D_k is uncorrelated with ϵ. As discussed in Section 2, students in some cases choose their roommates directly, and they generally have substantial influence over their dorm assignments either via direct room selection (such as for students in Learning Communities) or via their dorm preference list submitted to Housing. If individuals' dorm or roommate preferences are correlated with their academic abilities or preferences (which otherwise affect the outcome), the OLS estimate $\hat{\beta}_k^N$ will contain omitted variable bias and will not have a causal interpretation.

Commonly employed methods in the peer effects literature (such as Sacerdote (2001) and Carrell et al. (2013)) use stratified randomization designs, which involve estimation of equations such as (1) that are augmented with strata fixed effects that control for all nonrandom determinants of dorm/room/peer assignments. These equations may contain a large number of fixed effects, which can severely undermine efficient estimation of β_k even when all nonrandom determinants of assignments are known to the researcher. When the researcher knows the random determinants of treatment assignments (such as a randomly-generated tiebreaker) and the true mechanism used to assign treatments in addition to the aforementioned nonrandom determinants of assignments, it is possible to replace the high-dimensional strata fixed effects mentioned above with a scalar treatment propensity score calculated via simulation or via known analytic formulas as in Abdulkadiroğlu et al. (2017). Controlling for expected treatments in this way is sufficient to absorb the dependence between treatments and unobserved determinants of outcomes, allowing for a causal interpretation of $\hat{\beta}_k$ (Rosenbaum and Rubin (1983); Hirano and Imbens (2004)).

In our setting, we do not know the nonrandom determinants of assignments or the true mechanism used for treatment assignments because Housing exercises judgment in how they place students into rooms, potentially making use of student characteristics that we are unaware of.[3] We do, however, observe the random determinants of assignments in the form of Housing's randomly generated tiebreaker number assigned to each student. In this section, we describe a method that uses the random tiebreaker in conjuction with a researcher-specified assignment mechanism to recover treatment effect estimates with causal interpretations, which nests the above full-information approaches as special cases.

3.1 Room Choice

To facilitate the discussion of our method and related methods, we adapt notation from Abdulkadiroğlu et al. (2017) (AANP) to describe room allocations. A room choice problem assigns individuals $i = 1, 2, ..., N$ in set $\mathcal{I}$ to room-spots (beds) indexed by s, with $s = 1, 2, \ldots, S$. There are N total students and S total room spots, where a particular feature of our setting is that $N = S$ because the assignment of interest is a room-spot rather than a dorm.

A room assignment mechanism φ uses observed (to the researcher) nonrandom determinants of assignments, W_i, unobserved (to the researcher) nonrandom determinants of assignments, η_i, and randomly assigned tiebreaker, r_i, to place students into room-spots. The preferences that students have over dorms, as well as other traits relevant for room assignments (such as gender or athlete

[3]Less importantly, strata indicators for the determinants of assignments that we are aware of would involve fixed effects for 22! dorm preference permutations interacted with gender indicators, effectively using up our entire sample to estimate nuisance parameters.

status) are contained either in W_i or η_i depending on whether the researcher observes them. Each student has a type denoting nonrandom determinants of their room assignments used by mechanism φ indicated by $\theta_i^\varphi = (W_i^\varphi, \eta_i^\varphi)$, where $W_i^\varphi \subseteq W_i$ and $\eta_i^\varphi \subseteq \eta_i$.[4] Finally, each student is randomly assigned a tiebreaker number, r_i in $\{1, 2, ..., N\}$, which denotes their position in the order, and this tiebreaker is assumed to be observed by the researcher. A student's room-spot assignment as determined by mechanism φ is given by $\mu_i^\varphi(\theta^\varphi, r) \in \{1, 2, \ldots, S\}$, with i's realized assignment, μ_i, coinciding with the assignment determined by the true (potentially unknown to the researcher) allocation mechanism, denoted by $\varphi = *$, such that $\mu_i = \mu_i^*(\theta^*, r)$.

While we model assignments at the level of rooms, we estimate effects of lower-dimensional treatments that are deterministic functions of room assignments, where treatment k for individual i given by $D_{ik}^\varphi = D_{ik}(\mu^\varphi(\theta^\varphi, r))$ is potentially a function of all room assignments of all individuals. This allows us to estimate effects of general features of room assignments that are informative about policies outside of our immediate institutional context. In treatment k represents a dorm-level treatment assignment (such as inhabiting a highly gender-integrated dorm), the treatment is only a function of an individual's own room placement, so that we have $D_{ik}^\varphi = D_{ik}(\mu_i^\varphi(\theta^\varphi, r))$, where the individual's own assignment is still a function of all individuals types and random tiebreakers. Peer treatments are functions of both the individual's room placement and that of others (roommates and neighbors), which requires the more general notation $D_{ik} = D_{ik}(\mu^\varphi(\theta^\varphi, r))$ where treatment is a function of the entire vector of room assignments.

A contribution of this paper is to extend the setup of AANP to the case where we allow for arbitrary mechanism-specific determinants of assignments (θ^φ) for any mechanism that maps from these determinants to assignments, $\mu^\varphi(\theta^\varphi, r)$. As described by AANP, if all random and nonrandom determinants of assignments are observed, room assignments can be simulated for any mechanism that satisfies Equal Treatment of Equals (ETE), which requires that a mechanism gives all individuals of the same type the same assignment probabilities for all treatments. As discussed above, there is no guarantee in general that all determinants of assignments will be observed, and it seems unlikely that we observe all determinants of assignments in our application. We thus make a distinction between mechanisms that satisfy ETE generally and mechanisms that satisfy ETE with respect to observables for assignment to treatment k, which are defined as those that satisfy $\theta_i^\varphi \subseteq W_i$ and $\theta_i^\varphi = \theta_j^\varphi \implies \mathbb{E}[D_{ik}|\theta^\varphi] = \mathbb{E}[D_{jk}|\theta^\varphi]$.

Mechanisms that satisfy ETE with respect to observables for assignment to treatment k can be simulated to calculate φ-intended treatment assignments, D_{ik}^φ, as well as expected φ-intended treatment assignments, $\mathbb{E}[D_{ik}|\theta^\varphi]$, for treatment k. These simulated values are sufficient to construct conditionally exogeneous instruments for treatments, as we will discuss in the next subsection. Our distinction between mechanisms that satisfy ETE with respect to observables and those that do not allows, for instance, for cases in which a mechanism can be rightly said to satisfy equal treatment of equals (which has normative implications) without the researcher being able to simulate the mechanism, for instance due to data access limitations.[5] Furthermore, a mechanism may satisfy ETE with respect to observables for some treatments without doing so for other treatments. For instance, if students are assigned to dorms via Random Serial Dictator and are then sorted into rooms based on an unobserved trait (e.g. a preference for loud music known to Housing but not

[4]Related models in the school choice literature describe students' types in terms of their preferences over schools and the priorities they receive in schools. For a single mechanism, student traits such as gender can be modeled as priorities rooms give to particular students. Such priorities could vary between mechanisms, however, so we opt to model room preferences over students as part of the mechanism, which takes relevant student traits as inputs.

[5]For instance, if the true dorm assignment mechanism were Random Serial Dictator and Housing refused to provide us with student dorm preferences, the mechanism would satisfy ETE, but would not satisfy ETE with respect to observables.

to the researcher) then assignments under this mechanism satisfy equal treatment of equals with respect to observables for dorms but not for roommates. For such a mechanism, the methods we describe would permit estimation of dorm level treatment effects without permitting estimation of roommate level treatment effects.

3.2 Treatment Effect Estimation

Our empirical strategy is an intended intention to treat design, where we instrument for observed treatments with simulated values under an assumed mechanism (φ-intended treatments) while controlling for expected treatments implied by the same mechanism. In effect, we propose room assignments for students according to a partially random mechanism, Housing may or may not take up our proposed treatment assignments by passing them along to students, and students may or may not take up the assignments passed along to them by Housing. The econometric arguments are identical to those for standard intention to treatment instruments, where the takeup decision is an outcome of an unmodeled bivariate choice made by Housing with each student.

A general representation of the outcome equations we will estimate is

$$Y_i = \beta_0^{\varphi} + \beta_k^{\varphi} D_{ik}^{\varphi} + \beta^{\varphi} \mathbb{E}[D_{ik}^{\varphi} | \theta^{\varphi}] + \epsilon_i^{\varphi}, \tag{2}$$

where Y_i denotes an outcome of interest such as 4 year graduation, $D_{ik}^{\varphi} = D_{ik}(\mu^{\varphi}(\theta, r))$ is the φ-intended treatment k such as the share of i's roommates with different major STEM status, $\mathbb{E}[D_{ik}^{\varphi}|\theta^{\varphi}]$ is the expected value of the φ-intended treatment for student i, and ϵ_i^{φ} captures unmodeled determinants of outcomes. β_0^{φ} is a constant and the coefficient β_k^{φ} is the effect of φ-intended treatment k. Conditioning on the expected value of treatment (the propensity score for binary treatments) eliminates selection bias by siphoning out the variation in treatment that is nonrandom (Rosenbaum and Rubin, 1983; Hirano and Imbens, 2004).

Equation (2) allows consistent estimation of the effect of φ-intended treatment D_{ik}^{φ}, but our ultimate goal is estimation of the effect of realized treatment, D_{ik}. To establish our connection to the peer effects and school choice literatures, we begin with the special case of equation (2) that makes use of the true assignment mechanism,

$$Y_i = \beta_0^{*} + \beta_k D_{ik} + \beta^{*} \mathbb{E}[D_{ik}^{*} | \theta^{*}] + \epsilon_i^{*}, \tag{3}$$

where we have leveraged the definition of the true mechanism to impose $D_{ik}^{*} = D_{ik}$. School choice applications commonly estimate equations such as (3) by leveraging known mechanisms obtained via qualitative methods, such as interviews with school assignment adminsitrators.[6] In typical peer effects applications, the true assignment mechanism is unknown, which prohibits calculation (via simulation or otherwise) of $\mathbb{E}[D_{ik}^{*}|\theta^{*}]$. It is sufficient in such settings to proxy for $\mathbb{E}[D_{ik}^{*}|\theta^{*}]$ with a set of strata fixed effects $\delta_{\theta_i^{*}}$, yielding the specification

$$Y_i = \beta_0^{*} + \beta_k D_{ik} + \delta_{\theta_i^{*}} + \epsilon_i^{*}, \tag{4}$$

which permits consistent estimation of β_k if $\delta_{\theta_i^{*}}$ is properly specified (Rubin, 1977), at the cost of potentially losing degrees of freedom if there are a large number of nonrandom determinants of assignments.[7]

[6] School choice applications also frequently estimate intention to treat instrumental variables specifications, where the true assignment is known such that assignment administrators always take up $*$-intended treatment assignments, but students may not take up treatment assignments conveyed by adminitrators.

[7] A special case of assignment strata in the case of the peer effects literature is that individuals cannot be their own peer, which implies that individuals with each own-trait value for peer traits of interest (e.g. STEM intended major) face systematically different treatment probabilities. The common solution to control for individuals' own traits is sufficient to address this issue when fixed effects for values of individuals' own traits are interacted with other strata indicators.

In our application, the true mechanism is not known, so we are confined to outcome equations such as (2) that contain φ-intended and expected treatments for a proposed assignment mechanism. As described above, the effects of φ-intended treatments can be consistently estimated in this specification for an arbitrary mechanism. This parameter lacks immediate policy relevance, despite its causal interpretation. However, following the same arguments as we give above for the identification of the effect of D_{ik}^{φ} on Y_i, we can also identify the effect of D_{ik}^{φ} on D_{ik} from

$$D_{ik} = \gamma_0^{\varphi} + \gamma_k^{\varphi} D_{ik}^{\varphi} + \gamma^{\varphi} \mathbb{E}[D_{ik}^{\varphi}|\theta] + u_i^{\varphi}, \tag{5}$$

where γ_k^{φ} is the effect of φ-intended treatment k on realized treatment D_{ik}.% [8] The equations (2) and (5) together form the reduced form and first stage, respectively, of a just-identified instrumental variables model. It follows from the identification of β_k^{φ} and γ_D^{φ} that the effect of realized treatments on outcomes is also identified as

$$\beta_k = \frac{\beta_k^{\varphi}}{\gamma_k^{\varphi}},$$

under established instrumental variables assumptions.

3.3 Identification Assumptions

Because our empirical strategy uses intention to treat instrumental variables, we discuss our the validity of our IVs with respect to the sufficient conditions described by Angrist et al. (1996). We have concerns that some of these conditions do not hold in our setting, so we also discuss alternative assumptions as needed.

The stable unit treatment value assumption (SUTVA) requires that modeled treatment assignments mean the same thing for all individuals (Rubin, 1978, 1980, 1990). In the context of the school/dorm choice model with intended intention to treat instruments calculated from a misspecified assignment mechanism, this requires that each individual's modeled φ-intended treatment assignment and realized treatments are sufficient to define their treatments and outcomes. In our application, this means that if an individual responds to a φ-intended STEM major roommate by having a realized STEM roommate for some realization of the random tiebreaker vector (they take up the intended treatment), then they must always respond to φ-intended STEM roommate by having a realized STEM roommate. This is violated if individuals take up certain versions of intended treatments but not others, for instance if an individual is a complier for φ-intended Biology major roommate assignments but is a never-taker for φ-intended Engineering major roommate assignments. Similarly, SUTVA requires that outcomes are invariant to different varieties of modeled treatments, for example requiring that Biology major roommates have the same effects on outcomes as Engineering major roommates.[9] SUTVA violations complicate expression of treatment effects in terms of potential outcomes, and can contribute to inconsistent estimation of treatment effects by implying exclusion violations, which we will discuss below.

The assumption of random assignment of φ-intended treatment assignments is satisfied, conditional on expected φ-intended treatments, if the tiebreaker is truly random. If Housing altered certain individuals' tiebreakers after running an ETE assignment mechanism in order to ex-post rationalize placing certain students in certain rooms, this assumption would be violated. This assumption would similarly be violated if Housing "fished" for random number seeds that produce

[8] It is worth mentioning that the random tiebreaker, r_i, can be used as an instrument in place of D_{ik}^{φ}, with potential for a weaker first stage.

[9] The commonly-invoked example of treatment spillovers in a special case of treatment varieties, in which, for instance, having a STEM roommate who has STEM neighbors is a different variety of treatment than having a STEM roommate who does not have STEM neighbors.

room assignments that align with administrative objectives by rerunning the assignment mechanism multiple times. We consider Housing's admission that they overrule Random Serial Dictator assignments in accordance with their judgment to actually be evidence against either of the above violations of random assignment — if Housing wished for alternative assignments to those produced by a random procedure, they could make them without resorting to convoluted antics regarding the tiebreaker numbers. This assumption can be partially tested by checking whether obseved covariates predict tiebreaker values, which we show against for our setting in Appendix B.

The exclusion restriction requires that φ-intended treatment assignments only affect outcomes via realized peer or dorm traits. Three types of exclusion restriction violations come to mind. First, exclusion is violated if there are announcement effects in which assignment to treatments produce direct effects on outcomes, such as if preiminary interactions with a prospective STEM roommate during the summer effect outcomes regardless of whether the student takes up their assigned roommate in the fall. Second, exclusion is violated if the university uses tiebreaker numbers or intended treatments to inform non-housing treatment assignments, for instance if the university assigns course instructors or time-slots jointly with residences. We discussed data uses with adminstrative staff in the institution we study, and they emphasized that the residence hall data is not used for any non-housing purposes. We also note that in the institution we study, Housing data on dorm preferences, room assignments, and tiebreakers is kept in a separate data storage location with different data security staff and different data access procedures than other administrative data, which would make using this data for other purposes administratively difficult. Third, exclusion is violated if the modeled treatment (e.g. having a STEM major roommate) consists of multiple varieties of treatment with different treatment effects (e.g. having a Biology major roommate or an Engineering major roommate), which occurs with SUTVA violations.

For exclusion violations in general, estimated treatment effects retain policy relevance if the unmodeled channels (treatments) through which instruments drive outcomes in the sample are similar to those in the population. We assume that announcement effects in which mechanism-intended assignments affect outcomes separately from actual treatments are minimal and drive outcomes similarly in our institutional context as they do elsewhere, and we assume housing assignments and tiebreakers are not used for non-housing purposes.[10] Regarding different varieties of treatment, we assume our φ-intended treatment instruments satisfy uniform unordered monotonicity (Harris, 2022), which requires that instruments weakly increase the probability of receiving all unmodeled varieties of treatments for all individuals. This avoids situations such as the case where receiving a φ-intended STEM major roommate increases the probability of having a Biology major roommate but decreases the probability of having an Engineering major roommate, in which the unmodeled Engineering roommate effect receives negative weight in the modeled STEM roommate treatment effect estimand. Because of market clearing conditions (there are a set number of individuals with each major and a set number of room spots in each dorm), we do not expect our instruments to systematically shift individuals between different varieties of the same treatment, and we have no reason to expect such shifts within subpopulations with abnormal treatment effects. If individuals in our sample take up unmodeled varieties of treatment in response to intended assignment instruments proportinately to how individuals in the population take up varieties of treatment in response to true assignment mechanisms, our treatment effect estimates will not only weakly-

[10] It is implausible that mechanism-intended treatment assignments from mechanisms we specify have announcement effects if they do not correspond to summer assignments. Over 97% of students in our sample accept summer room offers (Table C.1), so effects of summer-assigned peers are approximately indistinguishable from effects of realized (fall term) peers. If other university settings have similar room offer takeup, then the technical exclusion violation here is a policy-relevant one, where our IV estimates will provide (ceteris paribus) valid predictions of peer effects in other settings.

positively weight effects of all varieties of treatment, but will place weights on effects of treatment varieties that correspond to the frequency of those treatment varieties in the population. Reference **?** instead of standard monotonicity.

A related issue involving unmodeled heterogeneity is that of heterogeneous treatment effects, which is typically addressed by the monotonicity assumption of Imbens and Angrist (1994). Monotonicity requires that simulated peer or dorm assignments induce individuals into corresponding realized treatment statuses without ever inducing any individuals out of those statuses. In our context, monotonicity implies that zero students have opposite-signed mismatch between their simulated peers and actual peers for all values of a simulated peer trait. This effectively rules out the possibility of random mistakes in our simulation of room assignments.[11] It also rules out the possibility that some students are discontented with their initial room (or roommate) assignment regardless of what it is, with the result being that they always request a new assignment before the start of the year. We expect such students to be rare, because of 97% of students take up their summer room offers, which we explore in Appendix C).

Instead of assuming strict monotonicity, we make the weaker alternative assumption that instruments that shift individuals into treatment on average do not shift subgroups of individuals with significantly different treatment effects out of treatment, or between varieties of a treatment. Within homogeneous treatments, this is the compliers-defiers condition of De Chaisemartin (2017). The net uniform unordered monotonicity condition of Harris (2022) extends the observed multiple treatment unordered monotonicity condition of Heckman and Pinto (2018) to the case with multiple unobserved versions of treatment, and allows for defiers with non-extreme treatment effects similarly to the compliers-defiers condition of De Chaisemartin (2017). We have no reason to expect our mechanism-intended treatment assignments to shift subgroups of individuals into or out of treatment systematically in accordance with their effects of treatments, because we restrict ourselves to mechanisms that make monotonic use of the random tiebreaker with respect to preferences — we impose that a higher value of r_i weakly increases the expected value of the preference rank of the assigned dorm and the room number in room order conditonal on receiving the same dorm.

To see why this restriction helps with monotonicity, consider the case where true assignments are done via Random Serial Dictator, and the researcher does not have access to student preferences over dorms. It is possible to infer the aggregate popularity of dorms from the random tiebreakers of their inhabitants, so it would possible to construct a common set of preferences for all students, and to construct simulated treatment assignments and expected simulated treatment assignments using Random Serial Dictator with the common set of preferences. Popular dorm treatment assignment instruments constructed from this procedure will generally predict realized assignments to popular dorms regardless of dorm preferences, because they encode the luckiness of each individual while ignoring their true preferences. It follows that individuals who prefer unpopular dorms will be defiers; if they get lucky random tiebreakers they will be erroneously assigned popular dorms, and if they get unlucky random tiebreakers they will be erroneously assigned unpopular dorms.[12] A selection-on-gains argument suggests that such individuals are likely to have lower treatment effects from assignment to popular dorms, so this sort of monotonicity violation is a nontrivial

[11] For instance, if a random 10% of students have actual peer assignments that differ from their simulated assignments (either because we make mistakes in matching the true assignment algorithm, or because administrative personel make mistakes in implementing it), we would expect 1% (10% of 10%) of students to be defiers for binary peer traits, because their simulated peers would differ from realized peers regardless of their simualted peer.

[12] The same problem arises when simply using the random tiebreaker as an instrument for treatment assignments directly, which is otherwise a valid strategy, though it is likely to produce weaker IVs than simulating reasonable mechanisms that make use of the room capacities of dorms and student preferences.

one. In simple terms, we assume that our simulated peer assignments will either match realized peer assignments with respect to both modeled and unmodeled peer traits, or that they will fail to match realized peer assignments *nonsystematically*.

The reelvance conditon requires that instruments meaningfully drive treatments. This requires that the mechanism proposed by the researcher bears some resemblance to that used by the assignment administrators. Any mechanism that leverages the random tiebreaker is likely to satisfy this condition in a strict sense, with the caveat that it may not if administrators don't actually use the random tiebreaker for assignments. Leveraging additional information correctly such as room/dorm capacities or student preferences should improve the strength of the instrument. It is possible to overfit the mechanism, with the extreme example being the case of using fixed effects for each (unique) value of the random tiebreaker as instruments, which collapses the IV model in (2) and (5) to a single linear regression of the outcome on the realized treatment. The next section discusses a data driven procedure for choosing a mechanism while avoiding problems associated with overfitting.

3.4 Instrument Selection

We begin the discussion of mechanism selection with a general representation of the problem, which will establish a connection to the model selection literature. The key insight is that there are an infinite number of potential mechanisms that we could propose for assigning students to dorm rooms. It follows that there are an infinite number of valid simulated instruments for peer and dorm assignments. These realizations lead us to instrumental variables model selection methods, such as those described by Belloni et al. (2012) and Belloni et al. (2014). Broadly speaking, these methods consider first stage equations that are similar to equation (5) that in principle allow for an arbitrarily large number of instrumental variables, such as

$$D_i = \gamma_0 + \sum_{\varphi=1}^{\Phi} (\gamma_D^{\varphi} D_i^{\varphi} + \gamma_E^{\varphi} \mathbb{E}[D_i^{\varphi} | \theta^{\varphi}]) + u_i, \tag{6}$$

with $\varphi = 1, 2, ..., \Phi$ indexing the proposed room assignment mechanisms, with the important point being that $\Phi >> N$.

Existing instrument selection methods search over many specifications like those in equation (6) and identify the one(s) that include the strongest instruments. While our problem is the same on a fundamental level, we face nonstandard challenges relative to common applications that use these methods for instrument selection. First, for us to construct an instrument, we must conceive of (or identify in the literature) a room assignment mechanism, code the mechanism in software, and run the code that generates instruments using the mechanism. This is costly in terms of cognition, human time, and computational time. Secondly, restricting ourselves to a single best-performing mechanism is particularly attractive for interpreting effects in the context of our institutional setting.[13] Finally, as we will discuss further in a moment, we are at particular risk of overfitting equation (6) if we search over an arbitrary number of mechanisms.

The particular risk of overfitting in our application comes from each student receiving a unique value of the tiebreaker, r_i, with simulated room-spot assignments being a unique and deterministic function of r and observables. The implication of this is that it is possible to construct simulated

[13] In other words, using simulated instruments from a single mechanism that is established in the mechanism design literature allows us to describe our reduced form effects from equations like (2) as "the effect of peers a person is assigned under mechanism φ on outcome Y_i". The inclusion of instruments from multiple mechanisms weakens this intuition in our view.

instruments and expected simulated treatments from proposed mechanisms that (1) satisfy Equal Treatment of Equals, (2) perfectly rationalize all observed assignments, yielding a deterministic first stage, and either (3a) explain all of the variation in actual treatments with simulated treatments, generating *seemingly* strong instruments or (3b) explain all of the variation in actual treatments with expected simulated treatments, generating weak instruments.

As an example of case (3a) above, consider the egregiously overfit mechanism $\varphi = **$ that satisfies $\mu = \mu^*(\theta^*, r) = \mu^{**}(r)$. This mechanism effectively observes the room-spot student i was assigned and infers that any student assigned random tie-breaker r_i in a counterfactual assignment allocation would receive room-spot μ_i, such that $\mu_j^{**}(r_i) = \mu_i^{**}(r_i)$ for all $j \neq i$. This mechanism trivially satisfies ETE because all students have equal probabilities of all room assignments, implying that all students of the same type do as well. Because $\mu_i(\theta, r) = \mu_i^{**}(r)$ for all i, it follows that the first stage in (5) collapses to

$$\begin{aligned} D_i =& \gamma_0^{**} + \gamma_D^{**} D_i^{**} + \gamma^{**}\mathbb{E}[D_i^{**}|\theta_i^{**}] + u_i^{**} \\ =& D_i^{**}. \end{aligned} \tag{7}$$

D_i^{**} has an unboudnedly large F-statistic regardless of sample size, with the expected value of simulated treatments playing no role in predicting actual treatment assignments. It is nonetheless an irrelevant instrument, because it does not systematically predict assignments for any draw of the random tiebreaker. There is minimal variation in $\mathbb{E}[D_i^{**}|\theta^{**}]$, so it will fail to capture unobserved determinants of room assignments in (2) as well. It follows that implementing a two-stage least squares regression using D_i^{**} as an instrument for D_i while controlling for $\mathbb{E}[D_i^{**}|\theta^{**}]$ is approximately equivalent to the naive OLS specification

$$Y_i = \beta_0 + \beta_k D_{ik} + e_i, \tag{8}$$

which is unlikely to identify causal estimates due to selection, as discussed above.[14]

As an example of case (3b) above where a deterministic mechanism produces weak instruments, consider the deterministic-on-observables mechanism $\varphi = \theta$ that satisfies $\mu = \mu^*(\theta, r) = \mu^\theta(\theta^\theta)$, where sufficient individual characteristics are contained in θ^θ to perfectly determine assignments. This mechanism effectively observes the room spot i was assigned and infers that they must have been nonrandomly assigned to that spot based on spot priorities over student characteristics, the simplest case of which is that preferences are ignored, and each room spots give special priority to the student that inhabits it. It follows that each individual would always be assigned their actual room in any counterfactual assignment allocation, with $\mu_i^\theta(\theta^\theta, r) = \mu_i^\theta(\theta^\theta, r')$ for all i for any alternative set of tiebreakers r'. In this case, the first stage in (5) collapses to

$$\begin{aligned} D_i =& \gamma_0^\theta + \gamma_D^\theta D_i^\theta + \gamma^\theta \mathbb{E}[D_i^\theta|\theta^\theta] + u_i^\theta \\ =& \mathbb{E}[D_i^\theta|\theta^\theta]. \end{aligned} \tag{9}$$

Here, D_i^θ is a weak instrument (with an F-statistic of zero regardless of sample size), with expected treatments completely explaining realized treatments. D_i^θ is perfectly collinear with $\mathbb{E}[D_i^\theta|\theta^\theta]$ so treatment effects in (2) are unidentified. We prefer mechanism θ to mechanism $**$ because it honestly reports its usefulness with an F-stat of zero, but both mechanisms are inadequate for treatment effect estimation.

[14] We say that instrumenting for D_i with D_i^{**} while controlling for $\mathbb{E}[D_i^{**}|\theta]$ is approximately equivalent to OLS with no controls because $\mathbb{E}[D_i^{**}|\theta]$ still accounts for individuals' inabilities to be their own peers. Each individual's value of $\mathbb{E}[D_i^{**}|\theta]$ will be constructed as the sample leave-one-out mean of D, which is similar to an OLS regression controlling only for individual i's own trait with no other room assignment strata indicators.

In order to identify a mechanism that avoids overfitting while also predicting treatment assignments, we place constraints on ourselves in our mechanism search. First, we interviewed University Housing prior to attempting to rationalize observed assignments with any mechanism, and we constrain ourselves to dorm assignment mechanisms that we discussed with them.[15] Second, we require that all candidate mechanisms make use only of characteristics that Housing told us they use: random tiebreakers, gender, nontraditional student status, year of study, athlete status, learning community status, predetermined roommate status, and dorm preferences. Third, we restricted ourselves to mechanisms that that are monotonic functions of random tiebreakers. Specifically, we require that the preference rank of the dorm a student is assigned to is monotonically increasing in the value of their random tiebreaker, ceteris paribus (early ranked students get their preferences over late students). Similarly, we also require that the position in mechanism-determined order of the room a student is assigned is monotonicially increasing in the value of their random tiebreaker conditional on dorm assignments (early ranked students are in "earlier" rooms, where earlier is defined by each mechanism, for instance, by a lower room number). Fourth, we do not explicitly target our peer or dorm treatments of interest, but instead we target all individual dorm assignments and peer random tiebreaker assignments. .

4 Data and Institutional Details

To estimate peer effects, we use data on undergraduate students who live in dorms from 2016 to 2019 at the University of Wisconsin-Madison, a large, public research university, with outcomes covering the same timespan. This administrative data contains information on the rooms students live in for years in which they live in residence halls, as well as a wealth of baseline variables such as gender, race, test scores, and intended majors. Additionally, it contains multiple imporant outcomes, including 4-year graduation rates (for the 2016 cohort), freshmen retention (for all but the 2019 cohort), graduation majors, and course grades (for all cohorts).Importantly, our data contains the random tiebreaker and residence hall preference lists provided by students, which are used to match them to their most-preferred dorms when possible.

Our empirical sample is formed of the subset of undergraduate students who live in dorm rooms during the fall term of their first year that are assigned the preceding summer. There are 24,265 undergraduate students who live in dorms in their first fall which are assigned during the preceding summer, of whom 17,972 have second year outcomes (retention and second year majors for those retained) and 5,902 have four year outcomes (graduation and graduation majors for those who graduate). We omit higher-year students from our sample even if they live in the dorms because they are relatively rare, we expect peer exposure to affect them less than first-year students, and because we observe many of them as first-years. Similarly, we omit first-year students who do not receive university residence hall room assignments the summer before their first fall term (for instance, due to submitting their dorm application late) because they are not included in the university's random room allocation process that is central to our empirical strategy.

The peer traits that we calculate are neighbor share female, roommate share STEM major distance, average roommate ACT math score, average roommate ACT verbal score, and distances between individuals own statuses and those of their roommates for ACT math and ACT verbal scores. For each of these, distances are calculated by taking the absolute value difference between an

[15]This restriction was incentive compatible, as conducting this interview was much easier than coding a room assignment mechanism.

individual and the peer trait of each relevant peer, then averaging over these distances.[16] Students are coded to have a STEM major if their intended major on their college application matches a major from the ICE list of STEM-designated majors (ICE, 2016). Students are coded as male or female based on their self-reported gender to university administration. ACT Math is defined by the score on the corresponding test, while ACT Verbal is the sum of the scores on the ACT Reading test and the ACT English test.[17] ACT scores come from administrative data, and we impute ACT scores for students with missing scores using official SAT to ACT conversion tables for students with SAT scores, and we impute scores using AP test count, AP test score average, high school rank, high school class size, and high school GPA for students with neither ACT scores nor SAT scores.

To estimate peer effects, it is necessary to identify students' peers. Students' roommates are readily identified as those individuals occupying the same room at the same time as a given student. To identify neighbors, we reference university blueprints and manually code x, y, and z coordinates for the doorway for each room on campus using drafting software. We then calculate Manhattan distances in inches between each room and every other room. This allows us to identify each room's nearest 3 rooms, the inhabitants of which are treated as neighbors for the student(s) in the room.

In addition to effects of exposure to various types of peers at the room level, we are also interested in effects of dorm-level assignments. We are particularly interested in characteristics of residence halls that are manipulable by university administrators, as these are most policy relevant at both the university we study and others. With this in mind, we estimate effects of dorms that are fully gender-integrated (men and women can live next door to one another), which we code as "coed" in tables. As we show in Table 1, about 70% of dorms at Unviersity of Wisconsin-Madison are fully gender-integrated, suggesting substantial scope for policy improvements if these dorms are found to have either positive or negative effects on student outcomes.

Our empirical strategy will control explicitly for the expected values of treatments, so additional controls are not necessary for identification of effects of interest. We will nonetheless include controls in our preferred specifications in the interest of increasing statistical precision. We control for gender, race, a nontraditional student indicator, a first generation college student indicator, ACT math score, ACT verbal score, and intended-major fixed effects (2-digit Classification of Instructional Programs codes). Summary statistics for the variables we use are available in Table 1.

5 Results

Our methods require selection of a mechanism to generate instrumental variables. We present model selection results in Section 5.1. We present treatment effect estimates of peers and dorms in Section 5.2, using instrumental variables generated from the chosen mechanism.

5.1 Mechanism Selection

We consider three dorm assignment mechanisms and three room assignment mechanisms that condition on dorms being assigned according to the dorm mechanism that best rationalizes realized dorm assignments. All of the mechanisms we consider have some common components. First, we hold fixed rooms for individuals described by Housing as not being randomly assigned: athletes,

[16] In words, we calculate the average distance between individual i and their peers, not the distance between individual i and their average peer.

[17] The SAT to ACT conversion table makes use of the same coding of the ACT Verbal score.

Table 1: Descriptive Statistics

	Mean (1)	SD (2)
First Term GPA	3.327	0.608
Freshmen Retention	0.957	0.204
4 Year Graduation	0.715	0.452
4 Year STEM Grad	0.292	0.455
Coed Dorm	0.697	0.460
Female	0.531	0.499
STEM	0.435	0.496
ACT Math	28.634	3.741
ACT Verbal	57.749	8.944
Nontraditional Student	0.016	0.125
In-State Student	0.556	0.497
First Generation Student	0.183	0.387
Asian	0.071	0.257
Black	0.018	0.132
Hispanic	0.056	0.230
White	0.707	0.455
Other Race	0.149	0.356
International	0.091	0.288
Fall 2016 Cohort	0.243	0.429
Fall 2017 Cohort	0.245	0.430
Fall 2018 Cohort	0.253	0.435
Fall 2019 Cohort	0.259	0.438
No Roommate	0.039	0.193
Random Room Assignment	0.691	0.462
Random Roommate Assignment	0.387	0.487
Observations	24265	

Notes: Means and standard deviations for outcomes, peer traits, and controls for all first year students in dorms..

nontraditional students, students in learning communities, and students with no recorded random tiebreaker. These individuals will not contribute to dorm effect estimates because their expected dorm is equal to their realized dorm. They may contribute to peer effect estimates because though they are not randomly assigned to rooms, their roommates and neighbors may be. Second, following advice from Housing, we place students in remaining dorms or room spots at random after all other students are assigned if their preferences are insufficient to place them according to the other rules of a mechanism. This is primarily relevant for individuals who do not provide complete preferences on their dorm preference sheet.

We consider three dorm assignment mechanisms. First, we consider random serial dictator, which assigns each individual to their preferred dorm if it is available, then to their next preferred, and so on, in order $r_i = 1, 2, ..., N$ without reference to other individuals. Next, we consider the Boston mechanism, which assigns individuals in order $r_i = 1, 2, ..., R$ to their 1st preferred dorm if possible, otherwise skipping them. Then it repeats this for unassigned individuals for their 2nd preferred dorm, and so on, until all students are assigned. Finally, because Housing reported occasionally deviating from RSD in an effort to fill unpopular dorms with students who like them *relatively* well, we consider a mechanism we call Boston→RSD which first assigns students to dorms via Boston *if and only if they are placed in an unpopular dorms*, otherwise skipping them. After all unpopular dorms are filled, this mechanism starts over and places remaining students in dorms according to RSD.[18]

We also consider three room assignment mechanisms, conditional on dorms being assigned according to the dorm mechanism that best rationalizes realized dorm assignments. The first, which we term Room ID order, is that students are placed in rooms in ascending room ID order as they are placed into a hall, where room ID is an administrative record that is distinct from room numbers. The second, which we term Room # order, is that students are placed in rooms in ascending room number order as they are placed into a hall, where room number is the publicly observable number for each room (for instance displayed next to the door). Finally we consider a Geographic order, which places students into rooms in a zig-zag using x and y coordinates for room doorways, with ascending z (floor) order.[19]

Standard tests of instrument strength would estimate equations such as (6) and consider the F-statistic associated with simulated instruments from each mechanism. We opt for an alternative approach due to the computational intensity of calculating $\mathbb{E}[D_{ik}(\mu^{\varphi}(\theta, r))|\theta]$ for all i and k for each mechanism. We estimate the following seemingly unrelated regression (SUR) model to evaluate mechanism accuracy for all J dorms in $j = 1, 2, ..., J$ for each dorm assignment mechanism,

$$\begin{aligned} Dorm_{i,1} &= \alpha^{\varphi}_{Dorm} Dorm^{\varphi}_{i,1} + u^{\varphi}_{i,1} \\ Dorm_{i,2} &= \alpha^{\varphi}_{Dorm} Dorm^{\varphi}_{i,2} + u^{\varphi}_{i,2} \\ &\vdots \\ Dorm_{i,J} &= \alpha^{\varphi}_{Dorm} Dorm^{\varphi}_{i,J} + u^{\varphi}_{i,J}, \end{aligned} \tag{10}$$

[18]The gist of this mechanism is that it will place people in unpopular dorms if they rank them relatively highly, regardless of their other dorm preferences. We suggested this mechanism to Housing after failing to match assignments with both RSD and Boston, and they told us it loosely approximates the sort of ad hoc deviations they occassionally make from the assignments implied by RSD. As shown in Table 2, it performs quite poorly.

[19]There are myriad plausible ways to code geographic room orders, all of which will produce very similar peer assignments if they respect our monotonic tiebreaker constraint. Generally, all rotations and reflections of assignments for symmetric dorms will produce identical peer assignments, while less extreme diversions (or approximations of such diversions in asymmetrical dorms) will produce similar peer assignments. We consider only a single geographic order rather than embarking on a specification hunt among many extremely similar mechanisms that explicitly replicate tiebreaker clustering geographically with room assignment clustering.

where we emphasize that the constant is constrained to 0 and all equations in the model are constrained to share a single slope coefficient. α^{φ}_{Dorm} has the attractive property of giving the weighted average over of all dorms of the probability of individual i having a realized assignment to dorm j conditional on having simulated assignment to dorm j. We similarly estimate accuracy measures for peer tiebreakers as

$$\begin{aligned} r_{i,0} &= \alpha^{\varphi}_{r} r^{\varphi}_{i,0} + \delta^{\varphi}_{i,0} + u^{\varphi}_{i,0} \\ r_{i,1-3} &= \alpha^{\varphi}_{r} r^{\varphi}_{i,1-3} + \delta^{\varphi}_{i,1-3} + u^{\varphi}_{i,1-3} \\ r_{i,4-7} &= \alpha^{\varphi}_{r} r^{\varphi}_{i,4-7} + \delta^{\varphi}_{i,4-7} + u^{\varphi}_{i,4-7}, \end{aligned} \tag{11}$$

where $r_{i,0}$ gives the average tiebreaker for individual i's realized roommate, $r_{i,1-3}$ gives the same for next door neighbors, $r_{i,4-7}$ gives the same for the next nearest neighbors. The φ superscript gives the same values for simulated assignments, with δ^{φ}_{i} denoting simulated dorm fixed effects. We report dorm assignment mechanism accuracy, α^{φ}_{Dorm}, in Panel 1 of Table 2 and room assignment mechanism accuracy, α^{φ}_{r}, in Panel 2 of Table 2 for all the mechanisms we consider.

Table 2: Mechanism Selection

	Random 1	Random 2	2016	2017-2019	Total
	(1)	(2)	(3)	(4)	(5)
Panel 1: Dorm Assignments					
RSD	0.865	0.878	0.961	0.852	0.872
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Boston	0.802	0.826	0.852	0.808	0.811
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
Boston → RSD	0.676	0.692	0.742	0.679	0.684
	(0.002)	(0.002)	(0.003)	(0.002)	(0.001)
Observations	8427	8427	3968	12886	16854
Panel 2: Peer Assignments					
Room ID	0.500	0.432	0.413	0.444	0.437
	(0.006)	(0.006)	(0.008)	(0.005)	(0.004)
Room #	0.515	0.445	0.425	0.459	0.451
	(0.006)	(0.006)	(0.008)	(0.005)	(0.004)
Geographic	0.511	0.437	0.426	0.450	0.445
	(0.006)	(0.006)	(0.008)	(0.005)	(0.004)
Observations	7659	7672	3705	11626	15331

Notes: Mechanism accuracy for dorms as measured by α^{φ}_{Dorm} from the seemingly unrelated regression model (10) in Panel 1 and mechanism accuracy for rooms as measured by α^{φ}_{r} from the seemingly unrelated regression model (11) in Panel 2. Mechanism details are described in the text. Robust standard errors of accuracy statistics in parentheses.

In the interest of thoroughness, we cross-validate mechanism accuracy for 5 subsets of students in our data who are randomly assigned to dorms and rooms (where we keep only those with roommates for the room assignment accuracy). The Random 1 and Random 2 samples are randomly chosen mutually exclusive halves of all randomly assigned students. We also consider 2016 (a year when University Housing staff reported they exercised less judgment in room assignments) separately from later years (when Housing staff exercised more judgment in room assignments). Finally, we also calculate accuracy for the entire sample of randomly-assigned students. We choose mechanisms with

the highest accuracy measures, RSD and Room #, for the total sample to construct instruments for inclusion in instrumental variables models described by equations (5) and (2), but we are encouraged by the cross-sample consistency in accuracy-maximizing mechanisms. We find that Random Serial Dictator is the best dorm assignment mechanism, with Room # order being the best room assignment mechanism — though all room assignment mechanisms have very similar accuracy. We note that we have smaller samples for accuracy measures because we only test accuracy for individuals who are randomly assigned to rooms, while the sample shrinks further for room accuracy checks because they rely on simulated and realized roommates, which are missing for individuals in single person rooms.

Our selected mechanisms produce simulated and expected dorm and peer treatment assignment statuses described in Table 3. There are two main takeaways from this table. First, our expected, simulated, and realized assignments have almost identical means for all variables.[20] Second, realized and simulated treatments have significantly more variance than expected treatments. Our identification strategy uses residual variation that is common to both simulated and realized treatments after conditioning on expected treatments to identify effects, so relatively low variance in expected treatments is good news.[21]

5.2 Treatment Effect Estimates

In this section we present treatment effect estimates for dorm and peer treatment assignments. In general, we estimate instrumental variables models that take the form

$$\begin{aligned} Y_i &= \beta_0 + \beta_k D_{ik} + \beta \mathbb{E}[D_{ik}^{\varphi} | \theta^{\varphi}] + \epsilon_i, \\ D_{ik} &= \gamma_0^{\varphi} + \gamma_k^{\varphi} D_{ik}^{\varphi} + \gamma^{\varphi} \mathbb{E}[D_{ik}^{\varphi} | \theta^{\varphi}] + u_i^{\varphi}. \end{aligned}$$

by two stage least squares, giving estimates of linear marginal effects of peer/dorm assignment k on outcome Y. We investigate effects of dorm coeducational status, neighbor gender, and roommate ACT scores and intended majors.

In general, we estimate models with a single treatment variable (or with an own trait interaction), rather than combining all treatments into a single model. If peer traits are correlated (they are), these specifications give easily interpretable policy relevant total effects of a particular peer or dorm trait on a particular outcome.[22] Such estimates are sufficient statistics for predicting the effect of counterfactual increases/decreases in assignments to a particular type or dorm or peer

[20] This is approximately true by construction as there is only one pool of students, though different realizations of assignments may differ in which students are placed in single-person rooms or rooms with more than two people, where their traits will receive less weight. Additionally, Housing sometimes changes the number of students in a room between summer assignments and fall assignments. Our realized treatments take room occupancy values values from fall assignments while our expected and simulated treatments take them from summer assignments. More extreme differences can occur with nonlinear peer traits, such as the peer trait distances measures (these could be set to zero for all students, in principle).

[21] A non-stratified RCT would produce approximately zero variance in expected treatments. In our application, expected treatments explain well over half of the variation in realized treatments, suggesting that most of the variation in peer and dorm assignments is driven by selection.

[22] Strictly speaking, these specifications are likely to produce exclusion restriction violations in the sense that the modeled trait is correlated with unmodeled traits that are contained in the error term. If the covariances between the modeled trait and unmodeled traits in our dataset are representative of those in the population, these total effect estimates retain external validity. This issue is ubiquitous in instrumental variables estimation. For instance, the effects of charter schools estimated by Abdulkadiroğlu et al. (2017) similarly are partially driven by in-sample covariances between charter school status and other unmodeled school characteristics (insofar as other characteristics have effects). One principled way to dispute findings from such studies (including ours) is to suggest that these in-sample covariances are significantly different in the research sample than they are in the population.

Table 3: Treatment Assignment Description

	Mean (1)	SD (2)
Coed Dorm	0.697	0.460
Simulated Coed Dorm	0.698	0.459
Expected Coed Dorm	0.698	0.370
Neighbor Female Share	0.521	0.368
Simulated Neighbor Female Share	0.518	0.371
Expected Neighbor Female Share	0.518	0.289
Roommate STEM	0.434	0.489
Simulated Roommate STEM	0.429	0.485
Expected Roommate STEM	0.430	0.382
Roommate STEM Distance	0.412	0.485
Simulated Roommate STEM Distance	0.424	0.484
Expected Roommate STEM Distance	0.423	0.381
Roommate ACT Math	28.650	3.676
Simulated Roommate ACT Math	28.647	3.648
Expected Roommate ACT Math	28.645	2.982
Roommate ACT Math Distance	3.643	2.857
Simulated Roommate ACT Math Distance	3.661	2.850
Expected Roommate ACT Math Distance	3.689	2.361
Roommate ACT Verbal	57.805	8.648
Simulated Roommate ACT Verbal	57.764	8.614
Expected Roommate ACT Verbal	57.760	6.877
Roommate ACT Verbal Distance	8.804	7.209
Simulated Roommate ACT Verbal Distance	8.877	7.189
Expected Roommate ACT Verbal Distance	8.920	5.958
No Roommate	0.039	0.193
No Simulated Roommate	0.047	0.211
Observations	24265	

Notes: Summary statistics for treatment variables as well as simulated treatments and expected simulated treatments from best-fitting mechanism (RSD assignments to dorms, Room # order assignments to rooms).

for a given outcome. Richer models with multiple treatments do not immediately inform likely effects of policies that change dorm/peer assignments with respect to single (for instance statistically significant) trait, because they estimate partial effects of assignment to many types of dorms and peers *conditional on other traits being held fixed*. In other words, our specifications can be readily used to answer questions of the form "what would happen if we assigned x more individuals to y type dorms/peers?", while specifications with multiple treatments cannot (additional information about covariances between peer and dorm traits is needed to use them for such predictions).

The question stated in the preceding paragraph is not policy relevant if relevant counterfactual policies are not feasible. In general, inelastically supplied treatments with homogeneous effects might not produce benefits by being redistributed, and the information needed to effectively redistribute them is not available if their effects are estimated in (misspecified) homogeneous effects models. Main effects of our treatments, therefore, are policy relevant to the situation in which dorms or peers with particular traits are not inelastically supplied. The implication of this for our application is that homogenous treatment effect specifications inform university policies that increase the availability of particular dorm or peer treatments, for instance by converting non-gender-integrated dorms into gender-integrated dorms, or by admitting more students with particular characteristics to the university (or assigning more of such students to university housing). In general, we consider such single effect specifications to be relevant for dorm effects, but to not be particularly relevant for peer effects.[23]

We show in Table 4 the effects of living in coeducational (highly gender-integrated) dorms on a range of academic outcomes. Of particular note, we find positive effects on four-year graduation rates (10.7 percentage points) and STEM graduation rates (14.6 percentage points) for men from inhabiting a coed dorm relative to a non-coed dorm. We find no evidence of positive or negative effects for women for four-year graduation, or for either men or women for other outcomes of interest. One policy implication from this result is that increasing the number of coeducational dorms on campus should be expected to increase male four-year (STEM) graduation rates without substantially affecting women. An alternative implication is that similar gains could be made by assigning all men to the existing coeducational dorms, while housing women in the less gender-integrated dorms as needed. While it is possible that the effects of coed dorms are unrelated to their gender compositions (which would raise concerns about external validity) or that they are nonlinearly related to their gender compositions, we are doubtful that such a counterfactual arrangement would preserve the mechanisms that are present within our sample through which coed dorms increase graduation rates for men.[24]

For peer effects specifications, we estimate models that include peer traits alongside own-trait interactions with peer traits. In these models, the main effect is of relatively little interest, while the interaction answers questions of the form "how much would outcome y change if we assigned a peer with trait x to an individual with that same trait instead of an individual without it?". For binary peer traits, the coefficient on the uninteracted term in these specifications gives the increase in an outcome for the group that lacks the trait. If there is a particular interest in increasing the outcome for this particular group (for instance to address preexisting inequalities), then these effects

[23] Even if a given university were to admit additional students with a particular trait (e.g. high verbal ACT scores) based on our findings or other insights, it is likely that any net benefits accrued by the university would be offset by negative effects on these students' second-most-preferred universities. It is possible that universities competing over students who produce desirable peer effects may improve market efficiency (for instance by increasing college attendance rates among such students due to increases in college-provided financial aid), but we expect these effects to be second order relative to benefits from selection-on-gains types of policies that leverage heterogeneous treatment effects to assign peers to students they will particularly benefit.

[24] The extreme case in which "coed" dorms are exclusively inhabited by men strikes us as a stark example of a policy that would be subject to the Lucas (1976) critique.

are policy-relevant. If only aggregate changes in outcomes are policy interest, then a significant (economically or statistically) effect for a particular group produces no relevant implications unless some other group has a different effect (with this information conveyed by the estimated effect of the interaction term).

We show effects of the share of neighbors that are female in Table 5. We find statistically insignificant point estimates that are consistent with our findings for coeducational dorms, suggesting that exposure to female neighbors may improve freshmen retention, and four-year graduation for men, with no offsetting negative effects for women (though these estimates are all statistically insignificant at conventional levels). We also find that female neighbors increase STEM majors among individuals who graduate, though these effects are not significantly different for men and women. The lack of statistical significance on differences in effects between men and women suggests limited scope to increase aggregate STEM graduation rates, but the statistically significant main effect does suggest that increases in gender segregation may reduce the STEM major gender gap (increasing STEM majors among graduating women and decreasing them among graduating men). We note that we have significantly less power to estimate effects of neighbors relative to dorms, as our preferred mechanism is approximately 87% accurate for dorm assignments, while being only 45% accurate for peer assignments (shown in Table 2). This is reflected in the substantially lower F-statistics for neighbor traits as compared to dorm assignments.

We next turn to effects of academic traits of roommates, with results shown in Table 6. We find that assignment to a STEM roommate statistically insignificantly reduces four-year graduation rates for non-STEM individuals, but that it has statistically-significantly smaller negative effects (by 8 percentage points) on STEM individuals. A statistically significant interaction effect in the opposite direction of a statistically insignificant main effect may seem to be of questionable policy relevance, but we reiterate our point above that the interaction effects are substantially more policy relevant than main effects because they are sufficient to predict aggregate effects of counterfactual peer assignments. Specifically, we would expect that assigning all STEM students to each other as peers would produce aggregate four-year graduation rates 3.1 percentage points higher than the alternative extreme case of assigning no STEM students to each other as peers, and 1.5 percentage points higher than the existing allocation, in which 52% of students with STEM intended majors have roommates with STEM intended majors. We find that receiving a peer with a standard deviation higher ACT math score has a statistically significant but economically insignificant differential effect on individuals with higher ACT math scores than those with low scores. We also find large and imprecise positive effects of assignment to roommates with high ACT verbal scores.

We have substantially less power to estimate effects of roommate traits, both because of our room assignment mechanism's poor accuracy for peer assignments relative to dorm assignments, and because only 39% of students in our sample are randomly assigned to roommates, approximately 10% of whom (4% of the total sample) receive no roommate due to being placed in a single-occupancy room. This is readily seen in our relatively low F-statistics in these models, which range from 71 for outcomes for which we observe a large portion of the sample to 6 for outcomes that we only observe for the 2016 incoming freshmen cohort. Our point estimates on ACT scores are in line with some past results such as those of Zimmerman (2003) which find relatively larger effects for verbal scores relative to math scores, though in our setting these effects are imprecisely estimated.

Interactions between own traits and peer traits allow for one form of policy-relevant heterogeneity in treatment effects. We also consider the distances between individuals' academic traits and those of their roommates. This allows us to directly answer questions regarding whether students should be clustered among peers of their own type or whether they should be exposed to peers who are different from them. For binary traits (such as gender and STEM major), an individual's peer

trait distance is a deterministic function of the peer trait of interest and the interaction between their own trait and the peer trait — implying that estimates of effects of peer trait distances will be similar to those obtained in the models with interactions described above. For traits where an individual's peers may have either higher or lower values of a trait than an individual, specifications that model effects of distances between traits convey unique information about potentially policy relevant nonlinearities in effects.

To see the potential value of individual-peer trait distance specifications relative to specifications with interactions, it is helpful to consider a DGP where the distance between an individual's trait and those of her peers drives outcomes for all types of individuals, but there are no linear effects for any group. To fix ideas for this example, we proceed with the simplifying assumption that each individual's expected peer trait is the same as their own trait, and we consider individuals at only two values of the trait (though their peers may have any value). Example data from this sort of model for an outcome Y and peer trait x_j is shown in Panel 1 of Figure 1. With such a DGP, it is tempting to conclude that the estimated effect of the own trait peer trait interaction in a model that includes this term alongside a peer trait main effect will yield a positive coefficient, as high trait individuals are particularly strongly affected by being assigned high trait peers (these matches have low trait distances). This ignores, however, that the identification strategy relies on controlling for expected peer traits. Controlling for expected peer traits produces a coefficient on the own-trait-peer-trait interaction that is driven by the comparison of outcomes between low trait individuals who are assigned to peers with higher or lower *than expected* traits and high trait individuals who are assigned to peers with higher or lower *than expected* traits. In other words, even though high trait individuals' peers have high traits more often than low trait individuals' peers, they do not have higher-than-expected traits more often. Controlling for expected peers in this scenario conditions out the variation in peer traits that would allow a negative peer trait distance effect to be captured by an own-trait-peer-trait interaction.[25]

The results on distances between individual's traits and their peer traits are shown in Table 7. We do not find statistically significant impacts for any trait considered. Interestingly, we obtain negative point estimates for roommate ACT Verbal score distance that are similar in magnitude to the positive main effects of ACT verbal scores in Table 6. Also, there are no significant differences in effects estimated in models that also include a linear term for average roommate ACT verbal scores.[26] Taking the narrowly insignificant effect of standard deviation increases in peer ACT verbal distance on freshmen retention at face value produces a prediction that assigning all students to roommates with identical verbal ACT scores (an approximately feasible policy) would increase freshmen retention by 8.8 percentage points, which would (implausibly) predict aggregate freshmen retention of 104.5%. The potential for large effects from feasible policy interventions regarding this trait is compelling, and we feel warrants future research on the role of peer verbal skill differences in determining outcomes, especially insofar as assignment to homogeneous groups has been previously found to increase academic achievement in classroom settings, for instance by Duflo et al. (2011) A potential explanation for such effects is that students communicate more easily with peers with similar verbal skills, and that effective communication between peers fosters incresased educational success and increases the amenity value of staying in school and facilitate coursework cooperation.

[25]The intuition that an interaction term will capture effects of distances between individuals' traits and those of their peers is correct for binary treatments for which all individuals have the same expected peer traits — an ex ante unlikely scenario that is at odds with the positive standard deviations on expected peer traits shown in Table 1.

[26]The minimal impact of also including main effects of peer traits alongside distances on estimates of distance effects is predictable. For students whose expected peer's trait is the same as their own trait, the correlation between peer trait distances and peer traits is mechanically driven toward zero when conditioning on expected traits — distant peers cannot have traits that are systematically higher than expected or lower than expected.

Table 4: Dorm Effects

	First Term GPA		Freshmen Retention		4 Year Graduation		4 Year STEM Grad	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Coed Dorm	-0.017	0.017	0.001	0.014	0.041	0.107*	0.048	0.146***
	(0.025)	(0.046)	(0.010)	(0.017)	(0.034)	(0.063)	(0.030)	(0.055)
Female × Coed Dorm		-0.056		-0.021		-0.104		-0.154**
		(0.054)		(0.021)		(0.074)		(0.065)
1st Stage F-stat	2115.449	366.931	1544.419	302.360	923.755	142.790	923.755	142.790
Mean of outcome	3.327	3.327	0.957	0.957	0.715	0.715	0.292	0.292
Observations	24265	24265	17972	17972	5902	5902	5902	5902

Notes: Effects of dorm assignments estimated via 2SLS instrumenting for realized assignments with simulated assignments. All specifications control for expected values of simulated instruments, controls listed in Table 1, and intended major fixed effects. Robust standard errors in parentheses. */**/*** denote significance at the 90, 95, and 99 percentage confidence levels.

Table 5: Effects of Peer Gender

	First Term GPA		Freshmen Retention		4 Year Graduation		4 Year STEM Grad	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
% Neighbor Female	0.041	0.062	0.018	0.028	0.061	0.111	0.190**	0.223*
	(0.071)	(0.112)	(0.028)	(0.039)	(0.086)	(0.134)	(0.080)	(0.128)
Female × % Neighbor Female		-0.044		-0.023		-0.088		-0.071
		(0.144)		(0.055)		(0.175)		(0.161)
1st Stage F-stat	474.531	236.057	329.546	75.264	198.845	54.876	198.845	54.876
Mean of outcome	3.327	3.327	0.957	0.957	0.715	0.715	0.292	0.292
Observations	24265	24265	17972	17972	5902	5902	5902	5902

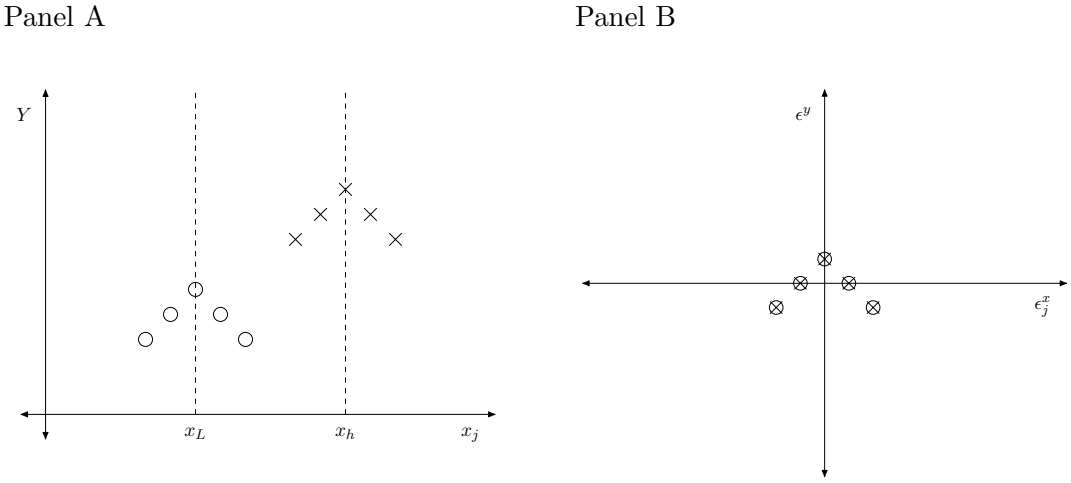
Notes: Effects of gender composition of next door neighbors estimated via 2SLS instrumenting for realized peers with simulated peers. All specifications control for expected values of simulated instruments, controls listed in Table 1, and intended major fixed effects. Robust standard errors in parentheses. */**/*** denote significance at the 90, 95, and 99 percentage confidence levels.

Table 6: Effects of Peer Academic Traits

	First Term GPA		Freshmen Retention		4 Year Graduation		4 Year STEM Grad	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Roommate STEM	-0.007	0.094	0.003	-0.135	-0.041	-0.202	0.054	0.168
	(0.175)	(0.299)	(0.075)	(0.141)	(0.243)	(0.362)	(0.231)	(0.319)
× STEM		-0.182		0.240		0.308		-0.221
		(0.366)		(0.165)		(0.487)		(0.462)
1st Stage F-stat	71.162	12.791	42.848	7.334	24.745	5.670	24.745	5.670
Roommate ACT Math	-0.124	-0.120	0.001	0.007	0.072	0.064	0.093	0.100
	(0.087)	(0.096)	(0.032)	(0.042)	(0.114)	(0.115)	(0.111)	(0.110)
× ACT Math		-0.014		-0.014		0.054		-0.051
		(0.102)		(0.039)		(0.096)		(0.089)
1st Stage F-stat	62.962	8.336	55.713	5.776	22.506	10.679	22.506	10.679
Roommate ACT Verbal	0.103	0.104	0.055	0.056	0.344*	0.327*	0.145	0.131
	(0.085)	(0.086)	(0.035)	(0.036)	(0.189)	(0.186)	(0.145)	(0.147)
× ACT Verbal		0.068		0.037		-0.108		-0.089
		(0.068)		(0.030)		(0.200)		(0.122)
1st Stage F-stat	52.549	26.055	40.797	18.540	12.795	5.699	12.795	5.699
Mean of outcome	3.330	3.330	0.957	0.957	0.718	0.718	0.290	0.290
Observations	22970	22970	16986	16986	5602	5602	5602	5602

Notes: Effects of peer academic traits of roommates estimated via 2SLS instrumenting for realized peers with simulated peers. All specifications control for expected values of simulated instruments, controls listed in Table 1, and intended major fixed effects. STEM is a binary indicator for having a STEM intended major, while ACT scores are normalized to have unit standard deviations. Robust standard errors in parentheses. Individuals with no roommates or no simulated roommates are dropped. */**/*** denote significance at the 90, 95, and 99 percentage confidence levels.

Figure 1: Example Data for Peer Trait Distances



Notes: Graphical representation of example data where outcomes are determined only by the individual's own trait and distances between an individual's trait and the same trait of their peers. Panel A shows a scatter plot of raw data for two types of individuals, with type L defined as $x_i = x_L$ shown as circles and type H defined as $x_i = x_H$ shown as x's, under the simplifying assumption that $\mathbb{E}[x_j|x_i] = x_i$ (each point could in principle represent any number of observations, with the caveat that the expected value for each group be unchanged). Panel B shows the Frisch-Waugh-Lovell residuals after controlling for the expected value of peer traits. It is apparent from Panel B that a linear regression of the outcome on peer traits, controlling for expected peer traits, will fail to uncover an economically significant relationship (regardless of interactions included in the model), while a distance-between-traits specification will succeed.

Table 7: Effects of Peer Academic Diversity

	First Term GPA		Freshmen Retention		4 Year Graduation		4 Year STEM Grad	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Roommate ACT Math Distance	0.054	0.043	0.005	0.005	0.034	0.030	0.127	0.119
	(0.097)	(0.102)	(0.035)	(0.037)	(0.132)	(0.134)	(0.129)	(0.131)
Roommate ACT Math		-0.117		0.002		0.073		0.094
		(0.090)		(0.035)		(0.114)		(0.112)
1st Stage F-stat	69.822	20.215	63.319	16.404	22.468	7.741	22.468	7.741
Roommate ACT Verbal Distance	-0.136	-0.114	-0.076	-0.061	-0.097	-0.168	-0.183	-0.215
	(0.108)	(0.108)	(0.047)	(0.049)	(0.226)	(0.296)	(0.205)	(0.239)
Roommate ACT Verbal		0.081		0.040		0.359*		0.160
		(0.086)		(0.037)		(0.207)		(0.166)
1st Stage F-stat	44.443	15.447	31.617	20.588	8.992	1.863	8.992	1.863
Mean of outcome	3.330	3.330	0.957	0.957	0.718	0.718	0.290	0.290
Observations	22970	22970	16986	16986	5602	5602	5602	5602

Notes: Effects of ACT score distances between roommates estimated via 2SLS instrumenting for realized peers with simulated peers. All specifications control for expected values of simulated instruments, controls listed in Table 1, and intended major fixed effects. ACT scores are normalized to have unit standard deviations, so distance effects give predictions of a peer with a one standard deviation higher or lower ACT score. Robust standard errors in parentheses. Individuals with no roommates or no simulated roommates are dropped. */**/*** denote significance at the 90, 95, and 99 percentage confidence levels.

6 Conclusion

This paper reports estimates of peer and dorm effects using a sample drawn from University of Wisconsin-Madison, a large public four-year university. We overcome substantial empirical challenges resulting from selection into dorms and peers by implementing an intent-to-intent-to-treat instrumental variables strategy that, vitally, is robust to misspecification of the unviersity's room assignment mechanism. Our methodological contribution is to extend methods commonly employed in the school choice literature to a setting where researcher-proposed treatment assignment mechanisms fail to replicate realized assignments. We perform a straightofrward model selection exercise which identifies a best-performing mechanism without requiring the computationally demanding simulation of expected treatments under alternative mechanisms. This method has potential to be of use in similar school (etc) choice settings in which researchers are unaware of assignment mechanisms used by central planners (or in which central planners do not explicitly use a consistent mechansim). More generally, our model selection procedure can be used to choose between alternative economic models which produce different predictions for realized and expected treatments in implementations of the instrumental variables methods described by Borusyak and Hull (2020).

Substantively, we find that highly gender-integrated dorms increase four-year graduation rates for men, while STEM peers reduce four-year graduation rates less for STEM individuals than for non-STEM individuals. We find promising suggestive evidence (falling short of conventional significance thresholds) that female neighbors increase freshmen retention for men and that high distance between ACT verbal scores reduces freshmen retention. Taken together, our findings suggest significant potential increases in freshmen retention and four-year graduation from approximately costless counterfactual room assignment mechanisms. Specifically, we predict that converting all dorms to be highly-gender integrated and exclusively assigning STEM students to each other as roommates would increase aggregate four-year graduation rates by approximately 3 percentage points relative to baseline.[27]

Our findings on gender effects complement a large literature that generally finds positive effects of girls and women on the educational outcomes of their peers. We acknowledge that our findings on the effects of coeducational dorms are identified off of the fourteen highly-gender-integrated dorms contained within our sample. With this small number of dorms, it is possible that an alternative amenity shared by coed dorms on campus is the true driver of outcomes. Future research estimating effects of gender-integration in dorms on outcomes at other universities would address this concern.

Our finding of aggregate benefits from clustering STEM individuals together as roommates complements some similar findings in the peer effects literature. For instance, Booij et al. (2017) find that group homogeneity of tutorial (study) groups improved performance among students at an economics and business school. Similarly, Duflo et al. (2011) find that students in Kenya benefitted from being in classroom environments with similar-ability peers. We broadly replicate these findings for effects of homogeneity of academic interests (STEM vs. non-STEM roommates) in addition to academic ability (distances between roommates ACT verbal scores), while also showing that they hold in residential settings in addition to classrooms and study groups. We expect these results to be particularly externally valid in other settings that use similar measures of STEM academic interests (and in which student's majors within STEM are similar) and verbal skills, such as other universities in the United States. We see no reason to expect substantial differences in effects for alternative measures of STEM interests or verbal skills, but studies at other universities could shed light on this. Along these lines, if students were to intentionally alter their verbal test scores

[27] Driven by a 30 percentage point increase in coed dorm assignments affecting 50% of the student population (males) by at a rate of 0.1, and a 48 percentage point increase in STEM roommate propensities affecting 43% of the student population (STEM individuals) at a rate of 0.07.

or (more plausibly) their reported intended majors to the university in an attempt to receive a particular peer assignment, this could invalidate the predictions from estimates. If universities do not publically (or even privately, as in the case of the university we study) reveal the details of their room assignment mechanisms, we expect that this sort of gaming by students will be unlikely.

We are particularly optimistic regarding the external validity of our findings. First, we estimate effects in a large public four-year university. Even if our findings are not generalizable to educational residential accommodations in other contexts (which they may be), we expect them to be generalizable to similar universities, which educate and house a large share of undergraduate students. Second, many of the counterfactual room assignment policies we discuss occur naturally within the support of our data. As discussed by Booij et al. (2017) and starkly investigated by Carrell et al. (2013), extrapolation outside of the support of data relies on potentially invalid functional form assumptions. Our estimates suggest potentially large effects from increases in the number of highly gender-integrated dorms as well as from clustering STEM students with each other as roommates. Even larger effects may be possible with clustering on verbal ACT scores, though are estimates are insufficient to predict effects of this sort of policy with certainty. Corroboration of these point estimates in other settings would potentially reduce this uncertainty, with substantial social benefits, though the consistency between our results and those of Duflo et al. (2011) and Booij et al. (2017) may be sufficient to motivate trials of such policies by university housing adminsitrators.

References

Abdulkadiroğlu, A., **Angrist, J.D.**, **Narita, Y.**, and **Pathak, P.A.** (2017). Research design meets market design: Using centralized assignment for impact evaluation. *Econometrica*, 85(5): 1373–1432

Angrist, J., **Gray-Lobe, G.**, **Idoux, C.M.**, and **Pathak, P.A.** (2022a). Still worth the trip? school busing effects in boston and new york. Working Paper 30308, National Bureau of Economic Research. doi:10.3386/w30308

Angrist, J., **Gray-Lobe, G.**, **Idoux, C.M.**, and **Pathak, P.A.** (2022b). Still worth the trip? school busing effects in boston and new york. Technical report, National Bureau of Economic Research

Angrist, J.D. (2014). The perils of peer effects. *Labour Economics*, 30: 98–108

Angrist, J.D., **Hull, P.D.**, **Pathak, P.A.**, and **Walters, C.R.** (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics*, 132(2): 871–919

Angrist, J.D., **Imbens, G.W.**, and **Rubin, D.B.** (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434): 444–455

Belloni, A., **Chen, D.**, **Chernozhukov, V.**, and **Hansen, C.** (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6): 2369–2429

Belloni, A., **Chernozhukov, V.**, and **Hansen, C.** (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2): 29–50

Booij, A.S., **Leuven, E.**, and **Oosterbeek, H.** (2017). Ability peer effects in university: Evidence from a randomized experiment. *The review of economic studies*, 84(2): 547–578

Borusyak, K. and **Hull, P.** (2020). Non-random exposure to exogenous shocks: Theory and applications. Technical report, National Bureau of Economic Research

Carrell, S.E., **Hoekstra, M.**, and **West, J.E.** (2019). The impact of college diversity on behavior toward minorities. *American Economic Journal: Economic Policy*, 11(4): 159–82

Carrell, S.E., **Sacerdote, B.I.**, and **West, J.E.** (2013). From natural variation to optimal policy? the importance of endogenous peer group formation. *Econometrica*, 81(3): 855–882

De Chaisemartin, C. (2017). Tolerating defiance? local average treatment effects without monotonicity. *Quantitative Economics*, 8(2): 367–396

Duflo, E., **Dupas, P.**, and **Kremer, M.** (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American economic review*, 101(5): 1739–74

Foster, G. (2006). It's not your peers, and it's not your friends: Some progress toward understanding the educational peer effect mechanism. *Journal of public Economics*, 90(8-9): 1455–1475

Gottfried, M.A. and **Harven, A.** (2015). The effect of having classmates with emotional and behavioral disorders and the protective nature of peer gender. *The Journal of Educational Research*, 108(1): 45–61

Harris, C. (2022). Interpreting instrumental variable estimands with unobserved treatment heterogeneity: The effects of college education. *Working Paper*

Heckman, J.J. and **Pinto, R.** (2018). Unordered monotonicity. *Econometrica*, 86(1): 1–35

Hill, A.J. (2017). The positive influence of female college students on their male peers. *Labour Economics*, 44: 151–160

Hirano, K. and **Imbens, G.W.** (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164: 73–84

ICE (2016). Stem designated degree program list. https://www.ice.gov/sites/default/files/documents/document/2016/stem-list.pdf, [online; accessed 2021-07-23].

Imbens, G.W. and **Angrist, J.D.** (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2): 467–475

Jones, T.R. and **Kofoed, M.S.** (2020). Do peers influence occupational preferences? evidence from randomly-assigned peer groups at west point. *Journal of Public Economics*, 184: 104154

Lavy, V. and **Schlosser, A.** (2011). Mechanisms and impacts of gender peer effects at school. *American Economic Journal: Applied Economics*, 3(2): 1–33

Lucas, R. (1976). Econometric policy evaluation: A critique. In *Theory, Policy, Institutions: Papers from the Carnegie-Rochester Conferences on Public Policy*, volume 1, 257. North Holland

Marmaros, D. and **Sacerdote, B.** (2006). How do friendships form? *The Quarterly Journal of Economics*, 121(1): 79–119

Rosenbaum, P.R. and **Rubin, D.B.** (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55

Rubin, D.B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1): 1–26

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 34–58

Rubin, D.B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371): 591–593

Rubin, D.B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4): 472–480

Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly journal of economics*, 116(2): 681–704

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of the Economics of Education*, volume 3, 249. Elsevier

Stinebrickner, R. and **Stinebrickner, T.R.** (2006). What can be learned about peer effects using college roommates? evidence from new survey data and students from disadvantaged backgrounds. *Journal of public Economics*, 90(8-9): 1435–1454

Zimmerman, D.J. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and statistics*, 85(1): 9–23

Appendix

A Instrument Validity

In this section we prove that instruments generated from a misspecified mechanism are sufficient to estimate LATEs within groups of individuals in which all randomly-assigned individuals could just as easily have received each others' treatment assignments. The intuition behind this result is that simulated instruments from misspecified mechanisms constitute guesses about treatment status, and while we may guess correctly or incorrectly for each person's treatment status, we will do so randomly. Individuals for whom we guess correctly will contribute to a positive first stage, while individuals for whom we guess incorrectly will contribute to a negative first stage. Treatment effects of the more common of these subgroups ("compliers") will be positively weighted by the IV estimand, while those of the less common subgroup ("defiers") will be negatively weighted, as established by Imbens and Angrist (1994). However, if all individuals could just as easily have received each others' assignments, then there will be no systematic differences in treatment effects between compliers and defiers– a condition we call *anonymous compliance*. It follows that the defiers' effects will be canceled out on average by those of compliers.

While, under our assumptions, compliance is independent of treatment assignment within each group this result does not hold between groups. For instance, we may guess individuals' treatment statuses correctly more often than incorrectly for most preference types, but vice versa for others. The latter preference types would thus be negatively-weighted on average by an IV that leveraged between-type comparisons. The takeaway of our result is that within-stratum instrument comparisons are valid, as well as positively-weighted averages of within-stratum comparisons. In the body of our paper, we reference this result for our preferred specifications, which interact recentered instruments with type indicators to allow for variation in first stage coefficients for different preference types.

A.1 Setup and Definitions

Here we provide notation and describe certain useful features of mechanisms that we use in our proofs. We extend theoretical results from Abdulkadiroğlu et al. (2017) (henceforth, AANP), so we adopt much of their notation.

We consider a market of N individuals indexed by i, where I gives the set of all individuals. There are S slots (rooms) indexed by $s = 1, ..., S$ to which individuals are assigned by a mechanism denoted by φ. Each slot has an enforced capacity under mechanism φ denoted by q_s^φ, with the capacities of all slots denoted by $\mathbf{q}^\varphi = (q_1^\varphi, ..., q_S^\varphi)$. We observe for each individual a vector of observed covariates W_i and a vector of L^φ realized lottery numbers $\mathbf{R}_i = [R_{i1}, ..., R_{iL^\varphi}]$, with $W = (W_i : i \in I)$ giving the matrix of observed covariates for all individuals and $\mathbf{R} = (\mathbf{R}_i : i \in I)$ giving lottery numbers for all individuals. We assume that for any $\ell \in \{1, ..., L^\varphi\}$, $R_{i\ell} = R_{j\ell}$ if and only if $i = j$.

Assignments are made using observed lottery numbers, observed characteristics for all individuals, and unobserved characteristics for all individuals. We define the distribution from which φ draws lottery numbers as h^φ, which has support $\mathcal{R}^\varphi$. Individual i has observed type $\theta_i^\varphi \subseteq W_i$ as defined by φ when making assignments, where i is said to be of type θ if $\theta_i^\varphi = \theta$. $\boldsymbol{\theta}^\varphi = (\theta_i^\varphi : i \in I)$ denotes the vector of observed types for all individuals, with Θ^φ denoting the set of all possible

observed types as defined by φ.[28] Similarly, individual i has unobserved type $\eta_i^\varphi \not\subseteq W_i$ as defined by φ when making assignments, where i is said to be of type η if $\eta_i^\varphi = \eta$. $\boldsymbol{\eta}^\varphi = (\eta_i^\varphi : i \in I)$ denotes the vector of unobserved types for all individuals, with $\mathcal{N}^\varphi$ denoting the set of all possible unobserved types as defined by φ. We assume that both θ^φ and η^φ are *fixed under re-randomization*, where we say that random variable X is fixed under re-randomization if $X_i(\mathbf{r}) = X_i(\mathbf{r}')$ for all $(\mathbf{r}, \mathbf{r}') \in \mathcal{R}^*$, for all i, where $\varphi = *$ denotes the true mechanism.[29]

A stochastic mechanism φ is a mapping from $(I, S, \mathbf{q}^\varphi, \Theta^\varphi, \mathcal{N}^\varphi)$ to a distribution of possible assignments. Its key components are a distribution from which it draws lottery numbers, defined above as h^φ, and an assignment function, ϕ^φ. ϕ^φ deterministically returns assignments for all individuals, taking $(I, S, \mathbf{q}^\varphi, \boldsymbol{\theta}^\varphi, \boldsymbol{\eta}^\varphi, \mathbf{R})$ as given. The assignment(s) for individual i produced by mechanism φ for a vector of lottery numbers $\mathbf{r}$ is given by $\phi_i^\varphi(\mathbf{r})$ $(1 \times M)$, which we assume is defined for all $\mathbf{r} \in \mathcal{R}^\varphi$, for all i, where we suppress the dependence of ϕ^φ on $(\mathbf{q}^\varphi, \boldsymbol{\theta}^\varphi, \boldsymbol{\eta}^\varphi)$, which we assume are fixed under re-randomization. The assignments for all individuals are then given by $\phi^\varphi(\mathbf{r}) = (\phi_i^\varphi(\mathbf{r}) : i \in I)$. Note that we allow for a single mechanism to make $M \geq 1$ assignments jointly (for instance, to college classes in different time slots in a term, such as in the application of **?**).

We also define a mechanism-invariant treatment function D, which give treatment statuses for all individuals as a function of assignments and covariates. The treatment status for individual i implied by mechanism φ given lottery realization $\mathbf{r}$ is then given by $D_i(\phi^\varphi(\mathbf{r}))$ where we suppress the potential dependence on W (as in the case of roommate assignments), and we sometimes use the shorthand $D_i^\varphi(\mathbf{r}) = D_i(\phi^\varphi(\mathbf{r}))$ to suppress the assignment function and denote the treatment as a mechanism-specific function of lottery numbers. We also denote the φ-intended treatment $D_i^\varphi = D_i(\phi^\varphi(\mathbf{R}))$, which is a random variable giving i's φ-intended treatment status as a function of the realized lottery number $\mathbf{R}$. D_i^* may be observed even if the true mechanism is unknown and/or $\mathbf{R}$ is unobserved, whereas D_i^φ can be simulated using $(\boldsymbol{\theta}^\varphi, \boldsymbol{\eta}^\varphi, \mathbf{R}^*, W)$, which requires that $\mathcal{N}^\varphi = \emptyset$ by the definition of η_i^φ as i's unobserved type.[30] Note that D_i is an individual-specific function of all assignments, such that treatment may be a function of other individuals' assignments, and individuals may or may not swap treatment statuses if they were to swap assignments.

We define some desirable properties of h^φ, ϕ^φ, and D, which will be instrumental in our proof below that intended treatments implied by mechanisms with such properties are valid instruments for realized treatments obtained from other mechanisms with such properties. In simple terms, we seek to define mechanisms under which observed subgroups of individuals approximately constitute strata in a randomized trial, which is to say members of each stratum are equally likely to receive treatment under either the true mechanism or a conjectured mechanism. In slightly less simple terms, we define mechanisms in which observed groups of individuals will reliably swap treatment statuses if they swap assignments, will reliably swap assignments if they swap lottery numbers, and have equal probabilities of all lottery numbers.

[28] AANP define types in terms of preferences and priorities over slots. None of our results strictly rely on this interpretation, though we expect that any general set of characteristics (such gender, grades, etc) can be reconfigured into the framework of AANP by defining preferences and priorities as functions of these characteristics.

[29] This condition implies the version stated by AANP, that X_i is independent of treatment given nonrandom determinants of assignments, while also imposing that it is independent of i's treatments implied by other mechanisms, their realized lottery, and the treatments, instruments, and lottery numbers of all other individuals. It follows that assuming potential outcomes are fixed under re-randomization, according to our definition, serves the same purpose as the stable unit treatment value and exclusion assumptions described in Angrist et al. (1996) and elsewhere.

[30] To emphasize, our methods could be used if treatment status were known, but assignments were unknown. Such a thing could occur if there is damage to the data, for instance if a data provisioner allowed researchers to observe treatment status but not assignment status, or if a dataset produced for a replication package contained treatments but not assignments.

Working from the inside out, we begin by discussing the lottery number generator, h^φ, which has support $\mathcal{R}^\varphi$. We assume that $\mathcal{R}^\varphi$ is of finite cardinality, referencing AANP's proof that this is without loss of generality. For some realization of lottery numbers, $\mathbf{r}$, let $(\mathbf{r}_{(i,j)}, \mathbf{r}_j, \mathbf{r}_i)$ denote the full set of lottery numbers when i and j swap numbers. We say that the distribution h^φ is *symmetric* if, for any i, j, and $\mathbf{r}$ such that $i \neq j$, and $\theta_i^\varphi = \theta_j^\varphi$,

$$h^\varphi(\mathbf{r}) = h^\varphi(\mathbf{r}_{(i,j)}, \mathbf{r}_j, \mathbf{r}_i).$$

Because the lottery number is the only determinant of assignments that is stochastic, we can define i's probability of receiving binary treatment D_i under mechanism φ as

$$P_i^\varphi = \sum_{\mathbf{r} \in \mathcal{R}^\varphi} h^\varphi(\mathbf{r}) D_i(\phi^\varphi(\mathbf{r})).$$

Next, we consider the assignment function. We define an assignment function as *partially-anonymous* if for any i, j, and $\mathbf{r}$ with $i \neq j$, $\theta_i^\varphi = \theta_j^\varphi$, and $\eta_i^\varphi = \eta_j^\varphi$,

$$\begin{aligned}
&\phi_i^\varphi(\mathbf{r}_{(i,j)}, \mathbf{r}_j, \mathbf{r}_i) = \phi_j^\varphi(\mathbf{r}), \\
&\phi_j^\varphi(\mathbf{r}_{(i,j)}, \mathbf{r}_j, \mathbf{r}_i) = \phi_i^\varphi(\mathbf{r}), \quad \text{and} \\
&\phi_k^\varphi(\mathbf{r}_{(i,j)}, \mathbf{r}_j, \mathbf{r}_i) = \phi_k^\varphi(\mathbf{r}) \quad \text{for all } k \in I \backslash \{i, j\},
\end{aligned}$$

which also describes mechanisms for which $\mathcal{N}^\varphi = \emptyset$, which we call *anonymous*. The partial anonymity property is helpful for describing fundamentally well-behaved mechanisms with annoying quirks that cause them to fall short of being strictly anonymous. For instance, it describes mechanisms that leverage information that may be unavailable to researchers (such as disability status in our application involving room assignments, or in applications involving military drafts), that otherwise approximate stratified randomized trials.

Finally, we discuss the treatment function, D. It is important to distinguish between treatments and assignments in our context because individuals with rare traits tend to have higher treatment propensities for roommate diversity than individuals with common traits, even when they have identical room assignment probabilities for all rooms. As we did with lottery numbers, let $(\phi_{(i,j)}^\varphi(\mathbf{r}), \phi_j^\varphi(\mathbf{r}), \phi_i^\varphi(\mathbf{r}))$ denote the full set of assignments when i and j swap assignments. We say that D is *stable* if for any φ, i, j, and $\mathbf{r}$ with $i \neq j$ and $\theta_i^\varphi = \theta_j^\varphi$,

$$\begin{aligned}
&D_i(\phi_{(i,j)}^\varphi(\mathbf{r}), \phi_j^\varphi(\mathbf{r}), \phi_i^\varphi(\mathbf{r})) = D_j(\phi^\varphi(\mathbf{r})), \\
&D_j(\phi_{(i,j)}^\varphi(\mathbf{r}), \phi_j^\varphi(\mathbf{r}), \phi_i^\varphi(\mathbf{r})) = D_i(\phi^\varphi(\mathbf{r})), \quad \text{and} \\
&D_k(\phi_{(i,j)}^\varphi(\mathbf{r}), \phi_j^\varphi(\mathbf{r}), \phi_i^\varphi(\mathbf{r})) = D_k(\phi^\varphi(\mathbf{r})) \quad \text{for all } k \in I \backslash \{i, j\}.
\end{aligned}$$

In other words, if individual i and j swap assignments, they also swap treatments.

The relevance of treatment function stability to peer effects is that individuals' own traits should be used to define the set of possible observed types Θ^φ, which recasts into a mechanism-design framework the insight of Angrist (2014) that mechanical relationships between own traits and peer traits are a threat to identification. Our notion of a *stable* treatment allows for simple cases in which treatment is a fixed mapping $D_i(\phi_i^\varphi(\mathbf{r}))$ for all i regardless of θ_i, such as charter school assignment as considered in AANP. It is also sufficiently general to cover cases in which treatment status is a function of other individuals' assignments $D_i(\phi^\varphi(\mathbf{r}))$, as in this paper's application to peer effects. Furthermore, it covers cases in which a given assignment implies different treatment statuses for different individuals, i.e. $D_i(\phi_i^\varphi(\mathbf{r})) \neq D_j(\phi_i^\varphi(\mathbf{r}))$ for some individuals i and j such that

$\theta_i \neq \theta_j$, such as the application in this paper concerning roommate diversity, or the application of Angrist et al. (2022b) in which the treatment is defined as the distance between i's residence and their assigned school.

In our proofs to follow, we make use of the notion of a *joint mechanism*. Joint mechanism $\varphi\varphi'$ is a mechanism that produces two sets of assignments for all individuals, which correspond to those produced by marginal mechanisms φ and φ' for any realization of lottery numbers. The joint mechanism $\varphi\varphi'$ draws lottery numbers from the distribution $h^{\varphi\varphi'}$, which has support $\mathcal{R}^{\varphi\varphi'}$. We assume that any draw from the joint distribution $h^{\varphi\varphi'}$ can be partitioned as $\mathbf{r} = (\mathbf{r}^{\varphi}, \mathbf{r}^{\varphi'})$, with $\phi^{\varphi\varphi'}(\mathbf{r}) = [\phi^{\varphi}(\mathbf{r}), \phi^{\varphi'}(\mathbf{r})] = [\phi^{\varphi}(\mathbf{r}^{\varphi}), \phi^{\varphi'}(\mathbf{r}^{\varphi'})]$ defined for all $\mathbf{r}$ in $\mathcal{R}^{\varphi\varphi'}$.

The definitions of anonymity, partial anonymity, and symmetry defined above for single mechanisms apply to a joint mechanism $\varphi\varphi'$ when defining i's observed type $\theta_i^{\varphi\varphi'}$ from the set $\Theta^{\varphi\varphi'} = \Theta^{\varphi} \times \Theta^{\varphi'}$ and unobserved type $\eta_i^{\varphi\varphi'}$ from the set $\mathcal{N}^{\varphi\varphi'} = \mathcal{N}^{\varphi} \times \mathcal{N}^{\varphi'}$. It is convenient but not necessary that $\Theta^{\varphi} = \Theta^{\varphi'}$ and $\mathcal{N}^{\varphi} = \mathcal{N}^{\varphi'}$. In essence, joint mechanisms using different information to produce assignments may require the researcher to specify more types of individual, but within each joint type, we will have an approximation of a randomized trial if the joint mechanism is well-behaved. We define i's joint distribution of intended treatments under joint mechanism $\varphi\varphi'$ as $g_i^{\varphi\varphi'}$ with

$$g_i^{\varphi\varphi'}(d, d') = \sum_{\mathbf{r} \in \mathcal{R}^{\varphi\varphi'}} h^{\varphi\varphi'}(\mathbf{r}) (1(D_i(\phi^{\varphi}(\mathbf{r})) = d) \times 1(D_i(\phi^{\varphi'}(\mathbf{r})) = d')),$$

giving the probability that $(D_i^{\varphi}, D_i^{\varphi'}) = (d, d')$ for any (d, d') such that d and d' are both in the range of D.

A.2 Proofs

In this section, we prove that the IV estimand conditional on observed type $\theta_i^* = \theta$ is a positively weighted average of treatment effects, under assumptions that we view as plausible in many central assignment settings. We begin by proving three lemmas which are used in our main theorem. The first describes sufficient conditions under which a mechanism satisfies *equal treatment of equals* (ETE), in which mechanism-intended treatments are equally likely for some subsets of the population. The second describes sufficient conditions under which *anonymous compliance holds*, a condition similar to ETE in which joint treatment assignments under multiple mechanisms are equally likely for subsets of the population. The third proves that differences in realized treatment shares between individuals with different values of conjectured-mechanism-intended treatments are driven by randomly-assigned individuals, if unobserved type η_i only affects assignments by rendering some individuals' treatments fixed under re-randomization.

We begin by showing that equal treatment of equals holds for treatments that are defined in other ways than in terms of i's own assignment, which extends the similar proof provided by AANP in their Appendix A.1.

Lemma 1: *If D is stable, ϕ^{φ} is anonymous, and h^{φ} is symmetric, then stochastic mechanism φ satisfies equal treatment of equals, that is*

$$P_i^{\varphi} = P_j^{\varphi}$$

for all i and j such that $\theta_i^{\varphi} = \theta_j^{\varphi}$.

Proof. Assume D is stable, ϕ^{φ} is anonymous, and h^{φ} is symmetric. Let i and j be such that $i \neq j$ and $\theta_i^{\varphi} = \theta_j^{\varphi}$. We assume without loss of generality (following AANP's proof) that $\mathcal{R}^{\varphi}$ has finite

cardinality. Let $\{\mathcal{R}, \mathcal{R}^{ij}\}$ be a partition of $\mathcal{R}^{\varphi}$ such that $\mathbf{r} \in \mathcal{R}$ if and only if $(\mathbf{r}_{(ij)}, \mathbf{r}_j, \mathbf{r}_i) \in \mathcal{R}^{ij}$. Beginning with the definition of i's probability of treatment under φ and using the shorthand $\mathbf{r}^{ji} = (\mathbf{r}_{(ij)}, \mathbf{r}_j, \mathbf{r}_i)$, we have

$$\begin{aligned}
P_i^{\varphi} &= \sum_{\mathbf{r} \in \mathcal{R}^{\varphi}} h^{\varphi}(\mathbf{r}) D_i(\phi^{\varphi}(\mathbf{r})) \\
&= \sum_{\mathbf{r} \in \mathcal{R}} h^{\varphi}(\mathbf{r}) D_i(\phi^{\varphi}(\mathbf{r})) + h^{\varphi}(\mathbf{r}^{ji}) D_i(\phi^{\varphi}(\mathbf{r}^{ji})) \\
&= \sum_{\mathbf{r} \in \mathcal{R}} h^{\varphi}(\mathbf{r}) D_j(\phi^{\varphi}_{(i,j)}(\mathbf{r}), \phi^{\varphi}_j(\mathbf{r}), \phi^{\varphi}_i(\mathbf{r})) + h^{\varphi}(\mathbf{r}^{ji}) D_j(\phi^{\varphi}_{(i,j)}(\mathbf{r}^{ji}), \phi^{\varphi}_j(\mathbf{r}^{ji}), \phi^{\varphi}_i(\mathbf{r}^{ji})) \\
&= \sum_{\mathbf{r} \in \mathcal{R}} h^{\varphi}(\mathbf{r}) D_j(\phi^{\varphi}(\mathbf{r}^{ji})) + h^{\varphi}(\mathbf{r}^{ji}) D_j(\phi^{\varphi}(\mathbf{r})) \\
&= \sum_{\mathbf{r} \in \mathcal{R}} h^{\varphi}(\mathbf{r}^{ji}) D_j(\phi^{\varphi}(\mathbf{r}^{ji})) + h^{\varphi}(\mathbf{r}) D_j(\phi^{\varphi}(\mathbf{r})) \\
&= \sum_{\mathbf{r} \in \mathcal{R}^{\varphi}} h^{\varphi}(\mathbf{r}) D_j(\phi^{\varphi}(\mathbf{r})) \\
&= P_j^{\varphi},
\end{aligned}$$

where the second line follows from the way the partition is defined, the third line follows from stability of D, the fourth line follows from ϕ^{φ} being anonymous if random, the fifth line follows from symmetry of h^{φ}, the sixth line follows from the way the partition is defined, and the last line follows from the definition of P_j^{φ}. □

Equal Treatment of Equals ensures that there are no systematic differences between individuals who receive treatment and those who do not, conditional on their type. The next lemma gives conditions that ensure, conditional on type, that there are no systematic differences between individuals who receive any combination of treatments under two distinct mechanisms and those who receive some other combination, a condition we call anonymous compliance.

Lemma 2: *If D is stable, ϕ^{φ} and $\phi^{\varphi'}$ are partially-anonymous, and $h^{\varphi\varphi'}$ is symmetric, then treatments determined by joint mechanism $\varphi\varphi'$ satisfy anonymous compliance with respect to each other, that is*

$$g_i^{\varphi\varphi'}(d, d') = g_j^{\varphi\varphi'}(d, d')$$

for all i and j such that $\theta_i^{\varphi\varphi'} = \theta_j^{\varphi\varphi'}$ and $\eta_i^{\varphi\varphi'} = \eta_j^{\varphi\varphi'}$, for all (d, d') such that d and d' are in the range of D.

Proof. Assume D is stable, ϕ^{φ} and $\phi^{\varphi'}$ are partially-anonymous, and $h^{\varphi\varphi'}$ is symmetric. Let i and j be such that $i \neq j$, $\theta_i^{\varphi\varphi'} = \theta_j^{\varphi\varphi'}$, and $\eta_i^{\varphi\varphi'} = \eta_j^{\varphi\varphi'}$. We assume without loss of generality (following AANP's proof) that $\mathcal{R}^{\varphi\varphi'}$ has finite cardinality. Let $\{R, R^{ij}\}$ be a partition of $\mathcal{R}^{\varphi\varphi'}$ such that $\mathbf{r} \in \mathcal{R}$ if and only if $(\mathbf{r}_{(ij)}, \mathbf{r}_j, \mathbf{r}_i) \in \mathcal{R}^{ij}$. Beginning with the definition of i's probability of receiving joint treatment (d, d') under joint mechanism $\varphi\varphi'$ and using the shorthand $\mathbf{r}^{ji} = (\mathbf{r}_{(ij)}, \mathbf{r}_j, \mathbf{r}_i)$, we

have

$$
\begin{aligned}
g_i^{\varphi\varphi'}(d,d') &= \sum_{\mathbf{r}\in\mathcal{R}^{\varphi\varphi'}} h^{\varphi\varphi'}(\mathbf{r})(1(D_i(\phi^{\varphi}(\mathbf{r}))=d)\times 1(D_i(\phi^{\varphi'}(\mathbf{r}))=d')) \\
&= \sum_{\mathbf{r}\in\mathcal{R}} h^{\varphi\varphi'}(\mathbf{r})(1(D_i(\phi^{\varphi}(\mathbf{r}))=d)\times 1(D_i(\phi^{\varphi'}(\mathbf{r}))=d')) \\
&\quad + h^{\varphi\varphi'}(\mathbf{r}^{ji})(1(D_i(\phi^{\varphi}(\mathbf{r}^{ji}))=d)\times 1(D_i(\phi^{\varphi'}(\mathbf{r}^{ji}))=d')) \\
&= \sum_{\mathbf{r}\in\mathcal{R}} h^{\varphi\varphi'}(\mathbf{r})(1(D_j(\phi^{\varphi}_{(i,j)}(\mathbf{r}),\phi^{\varphi}_j(\mathbf{r}),\phi^{\varphi}_i(\mathbf{r}))=d)\times 1(D_j(\phi^{\varphi'}_{(i,j)}(\mathbf{r}),\phi^{\varphi'}_j(\mathbf{r}),\phi^{\varphi'}_i(\mathbf{r}))=d)) \\
&\quad + h^{\varphi\varphi'}(\mathbf{r}^{ji})(1(D_j(\phi^{\varphi}_{(i,j)}(\mathbf{r}^{ji}),\phi^{\varphi}_j(\mathbf{r}^{ji}),\phi^{\varphi}_i(\mathbf{r}^{ji}))=d)\times 1(D_j(\phi^{\varphi}_{(i,j)}(\mathbf{r}^{ji}),\phi^{\varphi}_j(\mathbf{r}^{ji}),\phi^{\varphi}_i(\mathbf{r}^{ji}))=d')) \\
&= \sum_{\mathbf{r}\in\mathcal{R}} h^{\varphi\varphi'}(\mathbf{r})(1(D_j(\phi^{\varphi}(\mathbf{r}^{ji}))=d)\times 1(D_j(\phi^{\varphi'}(\mathbf{r}^{ji}))=d')) \\
&\quad + h^{\varphi\varphi'}(\mathbf{r}^{ji})(1(D_j(\phi^{\varphi}(\mathbf{r}))=d)\times 1(D_j(\phi^{\varphi'}(\mathbf{r}))=d')) \\
&= \sum_{\mathbf{r}\in\mathcal{R}} h^{\varphi\varphi'}(\mathbf{r}^{ji})(1(D_j(\phi^{\varphi}(\mathbf{r}^{ji}))=d)\times 1(D_j(\phi^{\varphi'}(\mathbf{r}^{ji}))=d')) \\
&\quad + h^{\varphi\varphi'}(\mathbf{r})(1(D_j(\phi^{\varphi}(\mathbf{r}))=d)\times 1(D_j(\phi^{\varphi'}(\mathbf{r}))=d')) \\
&= \sum_{\mathbf{r}\in\mathcal{R}^{\varphi\varphi'}} h^{\varphi\varphi'}(\mathbf{r})(1(D_j(\phi^{\varphi}(\mathbf{r}))=d)\times 1(D_j(\phi^{\varphi'}(\mathbf{r}))=d') \\
&= g_j^{\varphi\varphi'}(d,d'),
\end{aligned}
$$

where the second equality follows from the way the partition is defined, the third equality follows from stability of D, the fourth equality follows from ϕ^{φ} and $\phi^{\varphi'}$ being partially-anonymous, the fifth equality follows from symmetry of $h^{\varphi\varphi'}$, the sixth equality follows from the way the partition is defined, and the last equality follows from the definition of $g_j^{\varphi\varphi'}(d,d')$. Repeating the above for all combinations of (d,d') with d and d' in the range of D completes the proof. □

Next, we describe conditions under which differences in realized treatments for individuals with different values of the simulated instrument are driven entirely by individuals who are ex ante indistinguishable in their treatment propensity, if treatments are fixed under re-randomization for anyone whose unobserved type affects their assignments.

Lemma 3: *Let φ be a known mechanism used to construct instruments, let $*$ be the unknown true mechanism used to produce assignments, and let $*\varphi$ denote the joint mechanism of both $*$ and φ. If $h^{*\varphi}$ is symmetric, ϕ^{φ} is anonymous, ϕ^{*} is partially anonymous, D is stable, and D_i^* is fixed under re-randomization for all i such that $\eta_i^* \neq \eta^*$, it follows that the average difference in treatment propensities between individuals who do and do not receive the φ-intended treatment is identified as*

$$
\begin{aligned}
&P(D_i^*=1|\theta_i^{*\varphi}=\theta, D_i^{\varphi}=1) - P(D_i^*=1|\theta_i^{*\varphi}=\theta, D_i^{\varphi}=0) \\
&= P(\eta_i D_i^*=1|\theta_i^{*\varphi}=\theta, D_i^{\varphi}=1) - P(\eta_i D_i^*=1|\theta_i^{*\varphi}=\theta, D_i^{\varphi}=0)
\end{aligned}
$$

from the joint distribution of D^*, $\theta^{*\varphi}$, and $\mathbf{R}$.

Proof. Assume $h^{*\varphi}$ is symmetric, ϕ^{φ} is anonymous, ϕ^{*} is partially anonymous, D is stable, D_i^* is fixed under re-randomization for all i such that $\eta_i^* \neq \eta^*$, and that we observe $(D^*, \boldsymbol{\theta}^{*\varphi}, \mathbf{R})$. Because ϕ^{φ} is anonymous, the φ-intended instrument $D_i^{\varphi} = D_i(\phi^{\varphi}(\mathbf{R}))$ can be computed for all i, using the observed $(\boldsymbol{\theta}^{*\varphi}, \mathbf{R})$. Define $\eta_i = 1(\eta_i^* = \eta^*)$ as a binary indicator for random assignment under the true mechanism.

It follows that the "effect" of D_i^{φ} on D_i^* for type θ is identified as

$$\begin{aligned}
&P(D_i^* = 1|\theta_i^{*\varphi} = \theta, D_i^{\varphi} = 1) - P(D_i^* = 1|\theta_i^{*\varphi} = \theta, D_i^{\varphi} = 0) \\
&\quad = P(\eta_i D_i^* = 1|\theta_i^{*\varphi} = \theta, D_i^{\varphi} = 1) - P(\eta_i D_i^* = 1|\theta_i^{*\varphi} = \theta, D_i^{\varphi} = 0) \\
&+ P((1-\eta_i) D_i^* = 1|\theta_i^{*\varphi} = \theta, D_i^{\varphi} = 1) - P((1-\eta_i) D_i^* = 1|\theta_i^{*\varphi} = \theta, D_i^{\varphi} = 0) \\
&\quad = P(\eta_i D_i^* = 1|\theta_i^{*\varphi} = \theta, D_i^{\varphi} = 1) - P(\eta_i D_i^* = 1|\theta_i^{*\varphi} = \theta, D_i^{\varphi} = 0),
\end{aligned}$$

where the first equality follows from the substitution $D_i^* = \eta_i D_i^* + (1-\eta_i) D_i^*$, and the second equality follows because D_i^* is assumed to be fixed under re-randomization for all i such that $\eta_i = 0$ and φ satisfies equal treatment of equals, which completes the proof. □

Referencing Lemmas 1-3, we next define conditions under which misspecified mechanisms can be used to construct valid instruments.

Theorem 1: *Let φ be a known mechanism used to construct instruments, let $*$ be the unknown true mechanism used to produce assignments, and let $*\varphi$ denote the joint mechanism of both $*$ and φ. If $h^{*\varphi}$ is symmetric, ϕ^{φ} is anonymous, ϕ^* is partially anonymous, D is stable, $Y_i(1)$ and $Y_i(0)$ are fixed under re-randomization, and D_i^* is fixed under re-randomization for all i such that $\eta_i^* \neq \eta^*$, it follows that the conditional average treatment effect among the stratum defined by $(\theta_i^{*\varphi}, \eta_i^*) = (\theta, \eta^*)$, for any θ, is identified as*

$$\beta_\theta = \mathbb{E}[Y_i(1) - Y_i(0)|\theta_i^{*\varphi} = \theta, \eta_i^* = \eta^*]$$

from the joint distribution of Y, D^*, $\theta^{*\varphi}$, and $\mathbf{R}$, if $\mathbb{E}[Y_i|\theta_i^{*\varphi} = \theta, \tilde{\eta}_i = 1, D_i^{\varphi} = 1]$ and $\mathbb{E}[Y_i|\theta_i^{*\varphi} = \theta, \eta_i^* = \eta^*, D_i^{\varphi} = 0]$ are finite and $P(D_i^*|\theta_i^{*\varphi} = \theta, D_i^{\varphi} = 1) \neq P(D_i^*|\theta_i^{*\varphi} = \theta, D_i^{\varphi} = 0)$.

Proof. Assume $h^{*\varphi}$ is symmetric, ϕ^{φ} is anonymous, ϕ^* is partially anonymous, D is stable, $Y_i(1)$ and $Y_i(0)$ are fixed under re-randomization, D_i^* is fixed under re-randomization for all i such that $\eta_i^* \neq \eta^*$, and that we observe $(Y, D^*, \theta^{*\varphi}, \mathbf{R})$. Define $\eta_i = 1(\eta_i^* = \eta^*)$ as a binary indicator for random assignment under the true mechanism. By Lemma 1, it follows from the symmetry of $h^{*\varphi}$, anonymity of ϕ^{φ}, and stability of D that φ satisfies equal treatment of equals. By Lemma 2, it follows from the symmetry of $h^{*\varphi}$, anonymity of ϕ^{φ}, partial-anonymity of ϕ^*, and stability of D that joint mechanism $*\varphi$ satisfies anonymous compliance.

For any treatment function D, mechanism φ, and lottery number $\mathbf{r}$, $D_i(\phi^{\varphi}(\mathbf{r}))$ can be computed for all i if $\mathbf{r}$ is observed. It follows that the φ-intended instrument $D_i^{\varphi} = D_i(\phi^{\varphi}(\mathbf{R}))$ can be computed for all i using the realized lottery numbers $\mathbf{R}$ drawn from $h^{*\varphi}$ if $\mathbf{R}$ is observed. The φ-intended instrumental variable estimand for the effect of D_i^* on Y_i, conditional on $\theta_i^{*\varphi} = \theta$, is

$$\beta_\theta^{\varphi} = \frac{\mathbb{E}[Y_i|\theta_i^{*\varphi} = \theta, D_i^{\varphi} = 1] - \mathbb{E}[Y_i|\theta_i^{*\varphi} = \theta, D_i^{\varphi} = 0]}{P(D_i^* = 1|\theta_i^{*\varphi} = \theta, D_i^{\varphi} = 1) - P(D_i^* = 1|\theta_i^{*\varphi} = \theta, D_i^{\varphi} = 0)},$$

which can be computed if $(Y_i, D_i^*, \theta_i^{*\varphi})$ are observed in addition to the simulated D_i^{φ}.

Substituting potential outcomes for Y_i using $Y_i = Y_i(0) + D_i^*(\mathbf{R})(Y_i(1) - Y_i(0))$ and suppressing

the conditioning on $\theta_i^{*\varphi} = \theta$ for brevity, the numerator of the IV estimand is

$$\begin{aligned}
\alpha_\theta^\varphi &= \mathbb{E}[Y_i(0) + D_i^*(Y_i(1) - Y_i(0))|D_i^\varphi = 1] - \mathbb{E}[Y_i(0) + D_i^*(Y_i(1) - Y_i(0))|D_i^\varphi = 0] \\
&= \mathbb{E}[D_i^*(Y_i(1) - Y_i(0))|D_i^\varphi = 1] - \mathbb{E}[D_i^*(Y_i(1) - Y_i(0))|D_i^\varphi = 0] \\
&= \mathbb{E}[\eta_i D_i^*(Y_i(1) - Y_i(0))|D_i^\varphi = 1] - \mathbb{E}[\eta_i D_i^*(Y_i(1) - Y_i(0))|D_i^\varphi = 0] \\
&\quad + \mathbb{E}[(1-\eta_i) D_i^*(Y_i(1) - Y_i(0))|D_i^\varphi = 1] - \mathbb{E}[(1-\eta_i) D_i^*(Y_i(1) - Y_i(0))|D_i^\varphi = 0] \\
&\quad = \mathbb{E}[\eta_i D_i^*(Y_i(1) - Y_i(0))|D_i^\varphi = 1] - \mathbb{E}[\eta_i D_i^*(Y_i(1) - Y_i(0))|D_i^\varphi = 0] \\
&\quad\quad = P(\eta_i D_i^* = 1|D_i^\varphi = 1)\mathbb{E}[Y_i(1) - Y_i(0)|\eta_i^* = \eta^*, D_i^* = 1, D_i^\varphi = 1] \\
&\quad\quad\quad - P(\eta_i D_i^* = 1|D_i^\varphi = 0)\mathbb{E}[Y_i(1) - Y_i(0)|\eta_i^* = \eta^*, D_i^* = 1, D_i^\varphi = 0] \\
&\quad\quad\quad\quad = P(\eta_i D_i^* = 1|D_i^\varphi = 1)\mathbb{E}[Y_i(1) - Y_i(0)|\eta_i^* = \eta^*] \\
&\quad\quad\quad\quad\quad - P(\eta_i D_i^* = 1|D_i^\varphi = 0)\mathbb{E}[Y_i(1) - Y_i(0)|\eta_i^* = \eta^*] \\
&= (P(\eta_i D_i^* = 1|D_i^\varphi = 1) - P(\eta_i D_i^* = 1|D_i^\varphi = 0))\mathbb{E}[Y_i(1) - Y_i(0)|\eta_i^* = \eta^*].
\end{aligned}$$

The second equality follows from the linearity of the expectations operator, that φ satisfies ETE, and that $Y_i(0)$ is fixed under re-randomization.[31] The third equality is the result of the algebraic substitution $D_i^* = \eta_i D_i^* + (1-\eta_i)D_i^*$. The fourth equality follows from $Y_i(1)$, $Y_i(0)$, D_i^*, and η_i^* being fixed under re-randomization for individuals with $\tilde{\eta}_i = 0$, and that φ satisfies ETE. The fifth equality follows from η_i and D_i^* being binary and the definition of the conditional expectation. The sixth equality follows from the joint mechanism $*\varphi$ satisfying anonymous compliance and $Y_i(1)$ and $Y_i(0)$ being fixed under re-randomization. The final equality rearranges terms.

The instrumental variables estimand from above can thus be rewritten, explicitly conditioning on $\theta_i^{*\varphi} = \theta$, as

$$\beta_\theta^\varphi = \frac{P(\eta_i D_i^* = 1|\theta_i^{*\varphi} = \theta, D_i^\varphi = 1) - P(\eta_i D_i^* = 1|\theta_i^{*\varphi} = \theta, D_i^\varphi = 0)}{P(D_i^* = 1|\theta_i^{*\varphi} = \theta, D_i^\varphi = 1) - P(D_i^* = 1|\theta_i^{*\varphi} = \theta, D_i^\varphi = 0)}\mathbb{E}[Y_i(1) - Y_i(0)|\theta_i^{*\varphi} = \theta, \eta_i^* = \eta^*],$$

with Lemma 3 proving that the numerator and denominator of the fraction are equal, which completes the proof.

□

[31] ETE of φ implies that if $Y_i(0)$ is fixed under re-randomization, then $\mathbb{E}[Y_i(0)] = \mathbb{E}[Y_i(0)|D_i^\varphi = 1] = \mathbb{E}[Y_i(0)|D_i^\varphi = 0]$.

B Tiebreaker Randomization Check

The key to our research design is the randomization of room assignment tiebreakers, which are used in conjunction with student preferences to assign students to rooms. If tiebreakers are not actually random, our identification strategy is not valid. We note that Housing staff has told us that tiebreakers are random, while also telling us that their official policy is that staff judgment in assignments trumps any deference to the random allocation mechanism. It therefore seems to us that Housing has no incentive to doctor the random numbers to benefit some groups over others, even if they did have an interest in giving certain groups particularly advantageous room assignments.

To address the possibility of nonrandom tiebreaker numbers, we perform randomization checks by regressing random tiebreakers on the observed student characteristics described in Table 1. We report β coefficients from year-specific regressions of the form

$$r_i = X_i\beta + \epsilon_i,$$

where we test the null hypothesis that $\beta = 0$ for every element of X_i other than the constant.

Results are shown in Table B.1. Among 48 coefficients of interest, four are significant at the 95% confidence level and eight are significant at the 90% confidence level. At both levels of significance, we observe more statistically significant effects than would be expected with perfect randomization of tiebreakers (2.5 and 5). This is cause for concern. However, we note that the baseline characteristics that predict tiebreaker values do not have consistent signs, which suggests a lack of systematic bias

Table B.1: Initial Room Assignment Take-up by Student Type

	2016	2017	2018	2019
	(1)	(2)	(3)	(4)
Female	141.7*	-5.682	-59.01	22.71
	(79.01)	(80.35)	(77.17)	(76.90)
STEM	21.50	13.84	-32.14	-99.17
	(80.36)	(80.28)	(77.31)	(75.09)
ACT Math	29.50**	-8.882	-14.55	-2.817
	(12.56)	(12.93)	(11.75)	(11.33)
ACT Verbal	-5.628	8.257*	3.163	-8.251*
	(5.146)	(4.337)	(5.215)	(4.911)
Nontraditional	299.0	-185.2	76.96	640.0
	(312.2)	(275.5)	(319.2)	(489.3)
In-State	127.1	149.1*	-71.98	127.6
	(83.82)	(84.39)	(80.23)	(77.76)
First Generation	67.79	-18.86	-10.09	10.03
	(105.5)	(106.0)	(91.54)	(100.9)
Asian	-373.6*	282.0	375.5**	-342.4**
	(197.0)	(200.6)	(182.9)	(167.5)
Black	185.1	204.6	439.5	381.2
	(306.8)	(294.6)	(306.1)	(320.6)
Hispanic	88.53	97.36	270.3	-307.1
	(218.0)	(221.9)	(203.8)	(196.8)
White	134.8	90.98	314.7**	-206.4
	(150.8)	(152.7)	(137.5)	(135.5)
International	256.4	159.5	33.19	-152.2
	(191.6)	(201.1)	(175.6)	(175.5)
Constant	4119.7***	4474.3***	4904.9***	5515.5***
	(430.5)	(433.7)	(405.8)	(394.7)
Observations	5688	5778	5981	6269

Notes: Predictive associations between baseline student characteristics and tiebreakers, by year. *, **, and *** denote significance at the 90, 95, and 99% confidence levels.

C Room Assignment Monotonicity

As discussed in Section 3.3, a threat to identification is monotonicity violations for our simulated instrumental variables. Differences between simulated treatment assignments and actual treatment assignments arise either because some students are not assigned according to the mechanism we identify as best-fitting the data, and because some students are assigned in this way for their summer assignments, but request room changes for reasons such as conflict with roommates. The first group, students our assignment mechanism makes mistakes on, are likely individuals who either have unobserved room priorities in the rooms they receive (always takers or never takers), or they are random mistakes for whom the compliers-defiers assumption is likely particularly valid.

The second group of students who dislike their room(mate) assignment and request a new room are more problematic. It is possible to imagine a particularly cantankerous type of student who, upon being assigned a roommate of any type, takes issue with their roommate and requests a move. If such individuals have abnormal treatment effects from assignment to peers (for instance, if they are antisocial and have opposite signed effects from the general population for all peer traits), this will contribute to bias in treatment effect estimates if they are able to move to opposite-type peers when they request a room switch.

We have two arguments for monotonicity violations from cantankerous students being unlikely to invalidate our estimates. First, per interviews with Housing staff, students who request room moves are not allowed to request any particular alternative room(mate), they are placed near their old room in an open spot if one is available. It follows that they are not systematically likely to receive the opposite peer exposure from their initial assignment, except for the small mechanical effect from their prior room being unable to be their new room. It follows that even cantankerous students who reject initial assignments are often not defiers with respect to their peer or dorm treatment assignments.

Even for the subset of these students who are defiers, the compliers-defiers version of the monotonicity assumption requires only that there be sufficiently many compliers who share treatment effects with defiers to cancel them out. This means we need individuals in the complier population who relate to their peers similarly to the way defiers do. The initial assignment takeup rate is informative about the number of defiers there may be in our population, as we are primarily concerned about students rejecting their initial assignment and receiving a different realized assignment (we are not concerned about students having an initial assignment that doesn't match our simulated assignments). Statistics on room take-up are shown in Table C.1. We see that room assignment rejections are extremely rare at just over 2% of assignments. If every single individual who rejects their initial assignment has abnormal effects from exposure to peers, it is sufficient for us for the most similar individuals among the great mass of compliers to have overlapping treatment effects. Given that there appear to be at least 50 compliers for every defier, we feel confident that this overlap condition is satisfied.

Table C.1: Initial Room Assignment Take-up by Student Type

	Room Takeup Rate (1)
Total	0.976
Female	0.976
STEM	0.978
Above Median Math ACT	0.977
Above Median Verbal ACT	0.977
Nontraditional Student	0.958
In-State Student	0.979
First Generation Student	0.975
Asian	0.983
Black	0.967
Hispanic	0.971
White	0.976
Other Race	0.976
International	0.982
No Roommate	0.952
Random Room Assignment	0.978
Random Roommate Assignment	0.964
Observations	24265

Notes: Percentage of students of each type whose actual room is the same as their summer assignment.