Online Appendix

A Extensions

A.1 Relaxing Assumption 1

To clarify the role of Assumption 1, we can restate our hypotheses using more general notation:

$$\tilde{H}_0 : Y_i(Z) = Y_i(Z') \text{ for all } Z, Z' \text{ and for all } i \in \mathbb{U}$$

and

$$\tilde{H}_0^{\mathsf{w}_1,\mathsf{w}_2} : Y_i(Z) = Y_i(Z') \text{ for all } Z, Z' \text{ such that } w_i(Z), w_i(Z') \in \{\mathsf{w}_1, \mathsf{w}_2\} \text{ and for all } i \in \mathbb{U}.$$

If Assumption 1 holds, the null hypotheses $\tilde{H}_0$ and $\tilde{H}_0^{\mathsf{w}_1,\mathsf{w}_2}$ are equivalent to the null hypotheses H_0 and $H_0^{\mathsf{w}_1,\mathsf{w}_2}$; if it does not hold, the null hypotheses H_0 and $H_0^{\mathsf{w}_1,\mathsf{w}_2}$ are not well defined, while $\tilde{H}_0$ and $\tilde{H}_0^{\mathsf{w}_1,\mathsf{w}_2}$ can still be tested. In fact, the procedures in Section 3 used for testing H_0 and $H_0^{\mathsf{w}_1,\mathsf{w}_2}$ can be used without any modification to test $\tilde{H}_0$ and $\tilde{H}_0^{\mathsf{w}_1,\mathsf{w}_2}$ regardless of Assumption 1.

While Assumption 1 does not affect the mechanics of the test, it does impose restrictions on the alternative hypothesis, which changes the interpretation of rejecting the null hypothesis. In particular, Assumption 1 imposes two levels of exclusion restriction: one on the relevant attribute and one on the relevant group. Without this assumption, a number of different reasons could lead to rejecting the null hypotheses, H_0 or $H_0^{\mathsf{w}_1,\mathsf{w}_2}$. For instance, we would reject these hypotheses if a unit's outcome depends on the composition of attributes other than A, or if A is the relevant attribute but a unit's outcome depends on the composition of groups other than its own. Assumption 1 rules out both of these alternative channels for peer effects, narrowing the interpretation of rejecting the null hypotheses.

In summary, it is possible to test the null hypotheses $\tilde{H}_0$ and $\tilde{H}_0^{\mathsf{w}_1,\mathsf{w}_2}$ using the procedures in Section 3, regardless of the validity of Assumption 1. The price paid for the additional flexibility is that rejecting the null becomes less informative, since the alternative hypothesis includes channels of interference that were otherwise ruled out by Assumption 1.

As we discuss in the main text, there is little guidance for applied researchers on specifying exposure mappings, in part because these mappings can be highly context dependent. Thus, developing recommendations for exposure mappings in practice, as well as assessing sensitivity to those choices, is a necessary next step.

A.2 Testing weak null hypotheses

Our paper focuses on null hypotheses that impose a constant effect (usually zero) for all units. A natural question is how to extend our approach to *average* (or weak) null hypotheses. In

the no-interference setting, Wu and Ding (2020) propose permutation tests for weak null hypotheses using studentized test statistics. The result in Wu and Ding (2020, section 5.1) suggests that our permutation tests in Section 5 can also preserve the asymptotic type I error under weak null hypotheses with appropriately chosen test statistics. For example, we can test the following weak null hypothesis

$$H_0^{\mathsf{w}_1,\mathsf{w}_2} : \tau(\mathsf{w}_1, \mathsf{w}_2) = 0$$

where $\tau(\mathsf{w}_1, \mathsf{w}_2) = N^{-1}\sum_{i=1}^{N} Y_i(\mathsf{w}_1) - N^{-1}\sum_{i=1}^{N} Y_i(\mathsf{w}_2)$. Following the argument in Wu and Ding (2020), Procedure 2c will deliver an asymptotically valid p-value for $H_0^{\mathsf{w}_1,\mathsf{w}_2}$ if we use the studentized statistic

$$T(z; Y, \mathcal{U}) = \frac{\sum_{a\in\mathbb{A}} \pi_{[a]}(\hat{\bar{Y}}_{[a]\mathsf{w}_1} - \hat{\bar{Y}}_{[a]\mathsf{w}_2})}{\sqrt{\sum_{a\in\mathbb{A}} \pi_{[a]}^2(\hat{S}^2_{[a]\mathsf{w}_1}/n_{[a]\mathsf{w}_1} + \hat{S}^2_{[a]\mathsf{w}_2}/n_{[a]\mathsf{w}_2})}},$$

where $\pi_{[a]}$ is the proportion of $A_i = a$ among all units $i \in \mathbb{U}$, and $(n, \hat{\bar{Y}}, \hat{S}^2)$ are the sample size, mean and variance with subscripts denoting the attribute and exposure. As usual, we can also construct an asymptotic confidence interval for the average treatment effect $\tau(\mathsf{w}_1, \mathsf{w}_2)$ by inverting permutation tests.

A.3 Connection with the classic stratified, multi-arm trial

Our paper helps to clarify the relationship between randomized group formation experiments and traditional randomized stratified experiments in settings without interference or peer effects. In particular, we show that the designs we consider are equivalent to classic stratified randomized experiments with multiple arms. The non-sharp null hypotheses of interest correspond to contrasts between different arms of a multi-arm trial, possibly for a subset of units. Thus, at least with some reasonable simplifying assumptions, the otherwise complex setting of randomized group formation experiments reduces to a more familiar setup. As a byproduct, our proposed permutation tests are applicable to the classic designs as well.

B Additional analysis for Cai and Szeidl (2017)

This appendix section provides additional analysis and discussion of the re-analysis of Cai and Szeidl (2017) in Section 6.2.

B.1 Discussion of Assumption 1 — Alternative definitions of exposures

As discussed in Section 2.3, the interpretation of our test hinges on W being well-specified in the sense of Assumption 1. For instance, our tests could reject, in principle, even if H_0 was true but firm revenues differed across group assignments that produced the same peer

Table A1: Testing the sharp null under alternative exposures. 'one-sided' indicates the one-sided p-value (p) from Procedure 1b on a subpopulation; 'two-sided' is the corresponding two-sided p-value, $2\min(p, 1-p)$; '$*$' indicates a significant p-value at 5% level.

	$W_i^{(1)}$		$W_i^{(2)}$	
	one-sided	two-sided	one-sided	two-sided
small service firms	0.004*	0.007*	0.001*	0.002*
small manufacturing firms	0.980	0.041*	0.550	0.899
large service firms	0.607	0.785	0.262	0.523
large manufacturing firms	0.954	0.092	0.304	0.608

size exposure. Here, we explore the robustness of our results to two alternative specifications of the exposure. In the next section, we consider an additional specification, which reflects the type of peer group exposure that was actually randomized by Cai and Szeidl (2017).

In particular, we consider two additional definitions of exposures:

$$W_i^{(1)} = \frac{1}{|Z_i|}\sum_{j\in Z_i} \text{binary\_size}_j, \quad \text{or} \quad W_i^{(2)} = \frac{1}{|Z_i|}\sum_{j\in Z_i} \text{size}_j \cdot \text{revenue}_j,$$

where $\text{binary\_size}_j = 1$ if and only if firm j has size larger than the median size in j's region; and revenue_j is the log-revenue of firm j at baseline. The definitions capture coarser or finer versions, respectively, of our original exposure. For both these definitions, we run Procedure 1b and report the results in Table A1.

From Table A1, we observe that our results remain largely robust to the alternative exposure specifications we consider. For instance, across all specifications, we find a significant effect on small service firms, as in the previous section. There is one notable difference, however. Under the coarser exposure definition, $W_i^{(1)}$, we find evidence for a *negative* peer group effect on small manufacturing firms (two-sided p-value=0.04). This effect likely averages out the positive effect on small service firms (two-sided p-value=0.007), and produces a nonsignificant overall effect under $W_i^{(1)}$.

B.2 Pairwise null hypotheses

We now turn to pairwise non-sharp null hypotheses, extending the analysis of heterogeneity in the previous section. To that end, we focus on small manufacturing firms for which we observed a negative peer group effect in the previous section. We also consider a definition of treatment exposure that matches the type of exposure randomized in the actual experiment.

In particular, Cai and Szeidl (2017) randomized firms into 4 group types, namely, "small firms in the same sector", "large firms in the same sector", "mixed-size firms in the same sector", and "mixed-size firms with mixed sectors". We thus define the following discrete-

Table A2: Two-sided p-values and inverted randomization-based confidence intervals (at 5% level) for the pairwise weak nulls of Section B.2. 'n' indicates the number of units tested under the respective null, $H_0^{\mathsf{w}_1,\mathsf{w}_2}$; '$n_1$' is the number of firms exposed to w_1, and 'n_2 the number of firms exposed to w_2 ($n = n_1 + n_2$).

Null hypothesis	n (n_2/n_1)	p-value	point estimate	confidence interval
$H_0^{\text{S,SL}}$(small)	179 (84/95)	0.003	-0.449	(-1.062, -0.148)
$H_0^{\text{S,Sm}}$(small)	139 (44/95)	0.712	-0.549	(-1.084, 0.885)
$H_0^{\text{SL,SLm}}$(small)	188 (104/84)	0.903	0.017	(-0.445, 0.404)
$H_0^{\text{Sm,SLm}}$(small)	148 (104/44)	0.306	0.116	(-1.236, 0.387)

valued exposure for a small manufacturing firm i:

$$W_i^{(3)} = \begin{cases} \text{S}, & \text{if firm } i\text{'s peer group is all small manufacturing firms;} \\ \text{Sm}, & \text{if firm } i\text{'s peer group is all small firms of various sectors;} \\ \text{SL}, & \text{if firm } i\text{'s peer group is mixed-size manufacturing firms;} \\ \text{SLm}, & \text{if firm } i\text{'s peer group is mixed-size firms of various sectors.} \end{cases} \tag{B.1}$$

We consider four (weak) pairwise null hypotheses each comparing whether small manufacturing firms benefit from having a certain exposure level over another. For instance, $H_0^{\text{S,SL}}$(small) denotes a null hypothesis to test whether there are benefits of having a mix of large and small manufacturing peers as opposed to having only small manufacturing peers; $H_0^{\text{S,Sm}}$(small) denotes whether there are benefits of having a mix of small service or small manufacturing peers as opposed to having only small manufacturing peers; and so on.

Table A2 summarizes the results from using Procedure 2b on these pairwise null hypotheses. These results adds nuance to the negative peer group effect that we observed on small manufacturing firms in Table A1. In particular, we find that this negative peer group effect on small manufacturing firms is mainly due to their exposure to other large manufacturing firms. The relevant null, $H_0^{\text{S,SL}}$, is strongly rejected (two-sided p-value= 0.003), and the inverted confidence interval from this test indicates a range of 15% to 65% in revenue loss from such exposure. In contrast, no negative effects are observed when the exposure of small manufacturing firms is to small or large firms from a different sector (service).

C Simulation studies

This appendix section describes simulation studies that demonstrate the failure of asymptotic approximations in our applications and highlight the importance of using exact tests.

C.1 Simulation study calibrated to Li et al. (2019)

Our first simulation study illustrates the failure of asymptotics of the regression-based ("Neymanian") approach proposed by Li et al. (2019), in a setting calibrated to the roommates application in Section 6.1. Specifically, consider the following setup:

- $N = 156$ students allocated at random in rooms of size 4, indexed by i.
- A random $a\%$ of students (a is a free parameter) has $A = 1$ and the rest has $A = 0$.
- Sample $X_i \sim N(0, 1)$ iid; or $X_i = S_i\text{Weibull}(0.3)$, where S_i is random sign; or $X_i \sim$ mixture where mixture $= (1 - B)\delta_{-k} + BU[1 - \epsilon, 1 + \epsilon]$, where δ is the delta function, k, ϵ are constants and B is a Bernoulli random variable such that the mean is 0. All distributions are also normalized to have variance 1.
- Sample ε_i iid using the distributions described above.
- Define the exposure model, $W_i = \sum_{j \in \text{room}_i, j \neq i} A_j$, where room_i is the set of students in the same room as i.
- Define outcome $Y_i = 1 + 0 \cdot \mathbb{1}(W_i = 2) + X_i + (0.01 + A_i)\varepsilon_i$.

Note that, under this data-generating process, $Y^\omega(0) = Y^\omega(2)$ in distribution, and so our randomization tests remain finite-sample valid.

In this model, even though room allocation is completely randomized and there is no imbalance in room size, the joint distribution of (A, W) has a complex correlation structure due to the group formation design. In particular, roughly 3-5% of the units are exposed to $W = 2$, which results in a highly leveraged exposure assignment. Moreover, conditional on $W_i = 2$, unit i is more likely to be $A_i = 0$. Thus, under a mixture error distribution the outcomes Y_i of such units tend to be smaller than the outcomes under other exposures. This difference becomes negligible in the limit with more samples, but it is substantial in finite-samples, and cannot be easily captured by a regression model even under a robust specification.

To illustrate this point, we regress $Y_i \sim 1(W_i = 2) + X_i$ and use conservative heteroskedasticity-robust errors ("HC0"). We then test (at 5% level) the hypothesis that the regression coefficient of the exposure dummy variable is zero. A partial set of our results is shown in the table below. Here, we want only to show the pathological cases for the regression approach, and so we exclude the normal error setting for which regression performs well and near the nominal level.

Based on the results reported in Table A4, we observe that with Weibull errors (heavy tailed), the regression-based test has a size distortion and tends to under-reject. Under a mixture distribution for the errors, regression severely over-rejects. For instance, even with $N(0, 1)$ covariates, we observe rejection rates up to roughly 61%. In general, the regression-based test deteriorates under imbalanced designs.

In contrast, the randomization test is finite-sample valid as expected. Table A4 shows a

Table A3: Rejection rates from robust regression based on a simulation motivated by Li et al. (2019).

a ($\%A = 1$)	X	ε	Rejection rate%
10.00	$N(0,1)$	Weibull	1.63
30.00			1.20
50.00			1.40
10.00	Weibull	Weibull	1.54
30.00			1.50
50.00			1.50
10.00	mixture	Weibull	1.33
30.00			1.00
50.00			2.00
10.00	$N(0,1)$	mixture	61.98
30.00			11.40
50.00			10.10
10.00	Weibull	mixture	64.74
30.00			9.70
50.00			10.50
10.00	mixture	mixture	66.05
30.00			12.40
50.00			11.40

Table A4: Rejection rates (%) from robust regression and the group formation randomization test of Procedure 2b.

a ($\%A = 1$)	X	ε	regression	randomization test
10.00	$N(0,1)$	mixture	61.1	5.9
30.00			9.9	4.1
50.00			9.2	4.3

partial set of results relating to the pathological cases. We see that the randomization test achieves near-nominal level performance, with deviations from the nominal level due to Monte Carlo error.

C.2 Simulation study calibrated to Cai and Szeidl (2017)

We now consider the following simulation setup inspired by the analysis of Cai and Szeidl (2017) in Section 6.2. Here we focus on a subset of the data to illustrate the key intuition. We have 13 firms in the same sector and subregion, 2 of the firms are "large" and the remainder are "small." In particular, their sizes in terms of log number of employees are $A = (5, 5, Z_1, \dots Z_{11})$ where $Z_i \sim \text{Unif}[1, 3]$ are iid uniform. Following Cai and Szeidl (2017) we randomize the firms into two groups, one of type "mixed-size" (SL) and another of type "small-size" (S). Since Z_i are iid we can simply set as $L = (1, 1, 1, 2, 2, \dots, 2)$, such that group 1 is of type (SL) with two large firms and one small firm, and group 2 is of type (S) with all firms being small. The exposure of firm i is defined as the average group size of other firms in i's group:

$$W_i = \frac{1}{|\text{group}_i|} \sum_{j \in \text{group}_i} A_j.$$

We sample $\epsilon_i = N(0, \sigma_i^2)$ where $\sigma_i^2 = 1/|\text{group}_i|$ is the reciprocal of i's group size, and set the outcome model as $Y_i = 0 \cdot W_i + \epsilon_i$.

A conventional econometric approach would be to regress $Y \sim W + A$ and test whether the coefficient on W is zero, either through regular OLS errors or 'robust OLS'. However, both approaches are severely biased even when we condition on the same sector, subregion and firm sizes. In a simulated study with 10,000 replications based on this model, the nominal 5% rejection rate from regular OLS is 18.48%; and the rejection rate from robust OLS is 60.82%. For the same simulated data, the rejection rate of our randomization test is 4.8%.

The problem here is that OLS errors do not take into account the true correlation structure in W. For instance, in this model, both large firms have the exact same exposure regardless of the particular treatment assignment. Due to the problem structure, with high probability the errors in these two large groups can both be extreme leading to a spurious correlation between Y and W. Conditioning on firm characteristics in a regression model cannot fix this issue. In contrast, a randomization test can leverage the true correlation structure in W and has the correct level in finite-samples.

D Proofs

D.1 Proof of Theorem 1

Theorem 1. *Let $P(L)$ denote a distribution of the group labels with support $\mathbb{L} = \{1, \ldots, K\}^N$. Let $W = w^\ell(L) \in \mathbb{W}^N$ be the corresponding exposures, and let $U = u^\ell(L) \in \{0,1\}^N$ be the focal indicator vector, for some $w^\ell(\cdot), u^\ell(\cdot)$ defined by the analyst. Define $\mathbb{S}_{A,U} = \mathbb{S}_N(A) \cap \mathbb{S}_N(U)$, which is the permutation subgroup of $\mathbb{S}_N$ that leaves A (the attribute vector) and U (the focal unit vector) unchanged. Suppose that the following conditions hold.*

(a) $P(L) = P(\pi L)$, for all $\pi \in \mathbb{S}_{A,U}$ and $L \in \mathbb{L}$.

(b) $w^\ell(\cdot)$ is equivariant with respect to $\mathbb{S}_{A,U}$.

(c) $u^\ell(\cdot)$ is equivariant with respect to $\mathbb{S}_{A,U}$.

Then, W is uniformly distributed conditional on the event $\{W \in \mathcal{B}\}$, where $\mathcal{B} \in \mathcal{O}(\mathbb{W}^N; \mathbb{S}_{A,U})$.

Proof. We start with two lemmas.

Lemma D.1. *Suppose that Conditions (a)–(c) of Theorem 1 hold. Let $\mathcal{B} \in \mathcal{O}(\mathbb{W}^N; \mathbb{S}_{A,U})$ be an orbit such that $P(\mathcal{B}) > 0$. Then, for any $\pi \in \mathbb{S}_{A,U}$, we have*

$$P(\pi L \mid W \in \mathcal{B}, U) = P(L \mid W \in \mathcal{B}, U).$$

Proof of Lemma D.1. L determines both U and W, and so

$$P(W \in \mathcal{B}, U \mid L) = \mathbb{1}\{w^\ell(L) \in \mathcal{B}\} \cdot \mathbb{1}\{U = u^\ell(L)\}. \tag{D.1}$$

Similarly,

$$\begin{aligned}
P(W \in \mathcal{B}, U \mid \pi L) &= \mathbb{1}\{w^\ell(\pi L) \in \mathcal{B}\} \cdot \mathbb{1}\{U = u^\ell(\pi L)\} && \text{from (D.1)} \\
&= \mathbb{1}\{\pi w^\ell(L) \in \mathcal{B}\} \cdot \mathbb{1}\{U = \pi u^\ell(L)\} && \text{from Conditions (b)-(c)} \\
&= \mathbb{1}\{w^\ell(L) \in \mathcal{B}\} \cdot \mathbb{1}\{\pi^{-1} U = u^\ell(L)\} && \text{from orbit property of } \mathcal{B} \\
&= \mathbb{1}\{w^\ell(L) \in \mathcal{B}\} \cdot \mathbb{1}\{U = u^\ell(L)\} && \pi U = U \text{ since } \pi \in \mathbb{S}_{A,U} \\
&= P(W \in \mathcal{B}, U \mid L). && \text{from (D.1)}
\end{aligned} \tag{D.2}$$

It follows that

$$\begin{aligned}
& P(W \in \mathcal{B}, U \mid \pi L) P(\pi L) = P(W \in \mathcal{B}, U \mid L) P(L), && \text{From (D.2) and Condition (a)} \\
\Rightarrow & \frac{P(W \in \mathcal{B}, U \mid \pi L) P(\pi L)}{P(\mathcal{B})} = \frac{P(W \in \mathcal{B}, U \mid L) P(L)}{P(\mathcal{B})}, && \text{From } P(\mathcal{B}) > 0 \\
\Rightarrow & P(\pi L \mid W \in \mathcal{B}, U) = P(L \mid W \in \mathcal{B}, U).
\end{aligned}$$

□

Lemma D.1 shows that L retains its symmetry even conditionally on W beloning to some orbit $\mathcal{B}$ and conditional on focal selection U. The subspace where its symmetry holds is exactly the permutation subgroup $\mathbb{S}_{A,U}$, which leaves A and U fixed.

Lemma D.2. *Let $\mathsf{w} \in \mathbb{W}^N$ be a fixed exposure vector, and define*

$$\mathbb{L}(\mathsf{w}) = \{L \in \mathbb{L} : w^\ell(L) = \mathsf{w}\}.$$

Then, for any $\pi \in \mathbb{S}_{A,U}$, we have that

$$\mathbb{L}(\pi\mathsf{w}) = \{\pi L : L \in \mathbb{L}(\mathsf{w})\}.$$

Proof of Lemma D.2. The result follows from the equivariance property of w^ℓ in Condition (b). Specifically, equivariance implies that for any $L \in \mathbb{L}(\mathsf{w})$ then $\pi L \in \mathbb{L}(\pi\mathsf{w})$. Conversely, for any $L' \in \mathbb{L}(\pi\mathsf{w})$ then $\pi^{-1}L' \in \mathbb{L}(\mathsf{w})$. □

The crucial result in Lemma D.2 is that there exists a 1-1 mapping between the sets $\mathbb{L}(\mathsf{w})$ and $\mathbb{L}(\pi\mathsf{w})$ for any $\pi \in \mathbb{S}_{A,U}$.

We are now ready to prove the main result of Theorem 1. For a fixed $\mathsf{w} \in \mathbb{W}^N$:

$$P(W = \mathsf{w} \mid W \in \mathcal{B}, U) = \sum_{L \in \mathbb{L}} \mathbb{1}\{w^\ell(L) = \mathsf{w}\} P(L \mid W \in \mathcal{B}, U) = \sum_{L \in \mathbb{L}(\mathsf{w})} P(L \mid W \in \mathcal{B}, U). \tag{D.3}$$

Moreover, for any $\pi \in \mathbb{S}_{A,U}$:

$$\begin{aligned}
P(W = \pi\mathsf{w} \mid W \in \mathcal{B}, U) &= \sum_{L \in \mathbb{L}} \mathbb{1}\{w^\ell(L) = \pi\mathsf{w}\} P(L \mid W \in \mathcal{B}, U) && \text{From (D.3)} \\
&= \sum_{L \in \mathbb{L}(\pi\mathsf{w})} P(L \mid W \in \mathcal{B}, U) \\
&= \sum_{L \in \mathbb{L}(\mathsf{w})} P(\pi L \mid W \in \mathcal{B}, U) && \text{From Lemma D.2} \\
&= \sum_{L \in \mathbb{L}(\mathsf{w})} P(L \mid W \in \mathcal{B}, U) && \text{From Lemma D.1} \\
&= P(W = \mathsf{w} \mid W \in \mathcal{B}, U). && \text{(D.4)}
\end{aligned}$$

$\mathcal{B}$ is an orbit, and so it can be generated by any of its elements. Since $W \in \mathcal{B}$, the orbit can be generated by W, and so $\mathcal{B} = \{\pi W : \pi \in \mathbb{S}_{A,U}\}$. Therefore, conditional on $\{W \in \mathcal{B}\}$ and focals U, the orbit $\mathcal{B}$ is the entire domain of W. The result in (D.4) now implies that W is conditionally uniform given $\mathcal{B}$ and U. □

D.2 Proof of Lemma 1

Equivariance of w^ℓ. The exposure is defined in Eq. (3) as $w_i(Z) = \{A_j : j \in Z_i\}$. On the domain of group levels, this can be re-written as:

$$w_i^\ell(L) = \{A_j : L_j = L_i, j \neq i\}.$$

Now, let $\pi \in \mathbb{S}_N(A)$ be any transposition acting on L, i.e., a single swap between labels L_i, L_j of units i and j, respectively. After the swap, i is in the "room" that j was, and j is in the "room" that i was. From the definition of w^ℓ above, the exposures are only a function of other units' attributes in the room, and so units i and j swap exposures. The exposures of all units other than i, j are unaffected because i and j have the same attribute ($A_i = A_j$) due to $\pi \in \mathbb{S}_N(A)$.

Thus, we proved that $w^\ell(\pi L) = \pi w^\ell(L)$ whenever π is a transposition. Since every permutation is a composition of transpositions, the result holds for any permutation in $\mathbb{S}_N(A)$. Moreover, the result holds for $\pi \in \mathbb{S}_{A,U}$ as well since $\mathbb{S}_{A,U}$ is a subgroup of $\mathbb{S}_N(A)$.

Equivariance of u^ℓ. Recall the definition of focal selection in our setting, as defined in Eq. (9), $u_i(Z) = 1$ if and only if $w_i(Z) \in \{\mathsf{w}_1, \mathsf{w}_2\}$. With a slight abuse of notation, this can be re-written as $u^\ell(L) = \mathbb{1}\{w^\ell(L) \in \{\mathsf{w}_1, \mathsf{w}_2\}\}$, where the operation on the right-hand side is understood element-wise. Thus,

$$u^\ell(\pi L) = \mathbb{1}\{w^\ell(\pi L) \in \{\mathsf{w}_1, \mathsf{w}_2\}\} = \mathbb{1}\{\pi w^\ell(L) \in \{\mathsf{w}_1, \mathsf{w}_2\}\} = \pi \mathbb{1}\{w^\ell(L) \in \{\mathsf{w}_1, \mathsf{w}_2\}\}.$$

Here, the second equality follows from equivariance of w^ℓ and the last equality follows from the element-wise operation.

D.3 Proof of Lemma 2

In the stratified randomized design, define $\mathbf{m}^s : \mathbb{L}^N \to \mathbb{N}^{|\mathbb{A}| \times |\mathbb{L}|}$ as

$$\mathbf{m}^s(L)_{a,k} = \sum_{i \in \mathbb{U}} \mathbb{1}(L_i = k)\mathbb{1}(A_i = a),$$

which counts how many units with attribute $A_i = a$ are assigned to group label k. Then, a stratified randomized satisfies $P(L) \propto \mathbb{1}\{\mathbf{m}^s(L) = \mathbf{n}_A\}$, where $\mathbf{n}_A$ is fixed. For any

permutation $\pi \in \mathbb{S}_N(A)$, and any pair (a, k), we have

$$\begin{aligned}
\mathbf{m}^s(\pi L)_{a,k} &= \sum_{i \in \mathbb{U}} \mathbb{1}\{(\pi L)_i = k\}\mathbb{1}(A_i = a) \\
&= \sum_{i \in \mathbb{U}} \mathbb{1}(L_i = k)\mathbb{1}\{(\pi A)_i = a\} \qquad \text{From identity, } (\pi x)'y = x'(\pi y), \text{ for any } x, y \in \mathbb{R}^N \\
&= \sum_{i \in \mathbb{U}} \mathbb{1}(L_i = k)\mathbb{1}(A_i = a) \qquad \pi A = A \text{ since } \pi \in \mathbb{S}_N(A) \\
&= \mathbf{m}^s(L)_{a,k}. \qquad \text{(D.5)}
\end{aligned}$$

This results immediately implies that $P(\pi L) = P(L)$ for any $\pi \in \mathbb{S}_N(A)$. This holds also in the focal selection setting. That is, $P(\pi L) = P(L)$ for any $\pi \in \mathbb{S}_{A,U}$ since $\mathbb{S}_{A,U}$ is a subgroup of $\mathbb{S}_N(A)$. Thus, Condition (a) holds.

D.4 Proof of Lemma 3

In the completely randomized design, define $\mathbf{m}^c : \mathbb{L}^N \to \mathbb{N}^{|\mathbb{L}|}$ as

$$\mathbf{m}^c(L)_k = \sum_{i \in \mathbb{U}} \mathbb{1}(L_i = k),$$

which counts how many units are assigned to group label k. Then, $P(L) \propto \mathbb{1}\{\mathbf{m}^c(L) = \mathbf{n}\}$, where $\mathbf{n} = (n_1, \ldots, n_K)$ denotes how many units are to be assigned to each label, and is fixed. For any permutation $\pi \in \mathbb{S}_N$ and label k, we have

$$\mathbf{m}^c(\pi L)_k = \sum_{i \in \mathbb{U}} \mathbb{1}\{(\pi L)_i = k\} = \sum_{i \in \mathbb{U}} \mathbb{1}\{L_i = k\} = (L)_k. \qquad \text{(D.6)}$$

This results immediately implies that $P(\pi L) = P(L)$ for any $\pi \in \mathbb{S}_N$. This holds also for any subgroup of $\mathbb{S}_N$, including $\mathbb{S}_N(A)$ and $\mathbb{S}_{A,U}$. Both of these subgroups keep the attributes fixed, and so Procedures 1c and 2c in the completely randomized design are equivalent to the stratified randomized design with parameter $\mathbf{n}_A = \mathbf{m}^s(L)$. Thus, Condition (a) holds.