Robust Inference for Audit Studies: Simulation Methods and Results

Daniel S. Putman*

This Version: April 22, 2021
Click Here For Most Current Version

Contents

*Postdoctoral Fellow, Innovations for Poverty Action. dputman@poverty-action.org

List of Figures

List of Tables

1 Simulation Methods

1.1 Data Generating Processes

For each power simulation I start from a data generating process (DGP) that allows for correlation between transactions made by shoppers and agents. In these DGPs, i indexes auditor and j indexes target. For ease of exposition, I will refer to these as shoppers and agents, as is the case in the common example of mobile money overcharging.

1.1.1 Test Size Simulations

Let y_{ij} be a continuous outcome measure (e.g., charges per $ of cash out), for the size simulations, we model the DGP as

$$y_{ij} = \mu + \gamma_i + \delta_j + \varepsilon_{ij} \tag{1}$$

where γ_i captures shopper-specific shocks, and δ_j captures agent-specific shocks. In this case, $\mu = E(y_{ij})$. The shopper specific shock $\gamma_i \sim N(0, \sigma_S^2)$ where $\sigma_S^2 = \frac{\rho_S \sigma_\varepsilon^2}{1-\rho_S-\rho_A}$ and the agent specific shock $\delta_j \sim N(0, \sigma_A^2)$ where $\sigma_A^2 = \frac{\rho_A \sigma_\varepsilon^2}{1-\rho_S-\rho_A}$. For each case, I choose this expression to account for the multiple shocks built into the data structure, which causes the ICC for any given group to be reduced when using formulas, e.g., for one way intraclass dependence. By inflating the variance of any given shock by these factors, I able to match the empirical ICCs to ICC generated by simulation data. Finally, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ where $\sigma_\varepsilon^2 = (1 - \rho_S - \rho_A)Var(y_{ij})$.

1.1.2 Power: Shopper Effects

Shoppers i and j can belong to various groups G_l. Starting with a single shopper characteristic, shoppers i can belong to one of two groups: G_0, the comparison group, and G_1 the treatment group. We model the DGP as the following:

$$y_{ij} = \mu + \beta T_i + \gamma_i + \delta_j + \varepsilon_{ij} \tag{2}$$

where $T_i = \mathbf{1}(i \in G_1)$, $\mu = E(y_{ij}|T_i = 0)$, and β is the effect of being in the treated group relative to the comparison group.[1]

[1] We could similarly build a DGP if we have several groups of interest, $l = 0, \ldots, L$:

$$y_{ij} = \mu + \sum_{l=1}^{L} \beta_l T_{li} + \gamma_i + \delta_j + \varepsilon_{ij}$$

where $T_{li} = \mathbf{1}(i \in G_l)$ and β_l is the effect of being in a treatment group relative to the comparison group, though we don't use this in any simulations.

1.1.3 Power: Shopper and Agent Effects

I refer to "factorial" designs as a design where the treatment is the match between shopper and agent characteristics. One notable example is interactions between men and women as shoppers and agents. While these could take a few forms, a flexible form with one treatment group would be expressed

$$y_{ij} = \mu + \beta_1 T_i + \beta_2 T_j + \beta_3 T_i T_j + \gamma_i + \delta_j + \varepsilon_{ij} \tag{3}$$

where $T_i = \mathbf{1}(i \in G_1)$, $T_j = \mathbf{1}(j \in G_1)$ and β_1 is the effect of the shopper being in a treatment group, β_2 is the effect of the agent being in the treatment group, and β_3 the additional effect of both being in the treatment group (all relative to the comparison group). Here $\mu = E(y_{ij}|T_i = 0 \text{ and } T_j = 0)$.

1.2 Specifications

We can use audit studies to measure differential treatment by auditor characteristic, target characteristic, market characteristic, or some combination of the three. Mirroring the DGPs for power analysis, I focus on auditor characteristics, and then the combination of auditor and target characteristics. To fix ideas, I continue to use the example of discrimination that takes the form of differential charges by gender by mobile money agents.

1.2.1 Shopper Effects

In our example, we recruit shoppers that are 50% female and have them visit agents in their local market. Among the selected set of shoppers and agents, all shoppers visit all agents and attempt a transaction. They record how much they are charged for that transaction, which can be used to measure overcharging. After collecting this data, we run the following regression:

$$y_{ijm} = \alpha + \beta \, \mathrm{T}_{im} + \varepsilon_{ijm} \tag{4}$$

where i indexes shopper, j indexes agent, m indexes market, y_{ij} is a measure of overcharging, and T_{im} tracks if the shopper is female. α measures how much mean are (over)charged while $\alpha + \beta$ measures how much women are (over)charged. Then to test differential overcharging, we are interested in the hypothesis test $H_0 : \beta = 0$, which may be one or two sided.

1.2.2 Shopper and Agent Effects

We might also have a set up where the treatment of interest is the interaction of the shopper and the agent gender. In this case, y_{ijm} was the incidence of overcharging of auditor i by agent j.

However, treatment was defined based on the characteristics of both i and j. In particular, the paper tests for shopper-agent gender interaction effects, which we represent more simply as,

$$y_{ijm} = \alpha + \beta \mathrm{T}_{im} + \gamma \mathrm{T}_{jm} + \delta \mathrm{T}_{im}\mathrm{T}_{jm} + \varepsilon_{ijm}. \tag{5}$$

Here, α measures overcharging of men by men, $\alpha + \beta$ measures overcharging of women by men, $\alpha + \gamma$ measures overcharging of men by women, and finally, $\alpha + \beta + \delta$ measures overcharging of women by women. Here we are interested in three hypothesis tests: $H_0 : \beta = 0$, $H_0 : \gamma = 0$, and $H_0 : \delta = 0$.

1.3 Details of Simulations

1.3.1 Test Size Simulations

For these test size simulations, I vary ρ_S, ρ_A, n_S, and n_A. I assume three scenarios for shopper and agent ICCS: $(\rho_S, \rho_A) \in \{(0.05, 0.05),\ (0.20, 0.05),\ (0.05, 0.20)\}$, Likewise, I assume three scenarios for number of shoppers and agents: $(n_S, n_A) \in \{(4, 16),\ (8, 8),\ (16, 4)\}$. In total, nine different parameter sets are used, combining all possible tuples of (ρ_S, ρ_A) and (n_S, n_A) for these values. All other parameters in the simulation are held fixed. In particular, I assume $\sigma^2_\varepsilon = 1$, $\rho_M = 0$, and $n_M = 30$.[2] Shoppers and agents are both assumed to be 50% female and 50% male in each market. It is assumed each shopper makes one transaction with each agent within a market, for a total of 64 transactions in each market for all scenarios. Thus for all simulations, $64 \times 30 = 1920$ transactions take place.

For each simulation, I plot the standard errors for IID, One-Way: Agent, One-Way: Shopper, Two-Way, and One-Way Market clustered standard errors. Additionally, I evaluate the rate at which we reject the null hypotheses for both the shopper only specification and the factorial specification and present these results.

2 Test Size Simulation Results

2.1 Standard Errors

2.1.1 Shopper ICC Equals Agent ICC

Figures 1, 2, and 3 show standard errors when shopper ICC equals agent ICC, $\rho_S = \rho_A = 0.05$. For the shopper specification (top left), one-way agent and IID errors are relatively small compared

[2] Only three of ρ_S, ρ_A, σ^2_ε and $Var(y_{ij})$ need be specified, so $Var(y_{ij})$ is not specified, though depending on ρ_S and ρ_A, it should be $\frac{1}{1-\rho_S-\rho_A}$

to two-way and one-way shopper errors. With few markets, standard errors clustered one-way on market don't converge as quickly as the others here, which persists across simulations and specifications.

When considering the factorial specification, these results are consistent for the shopper effect (top right). However, as would be expected, when considering the agent effect (bottom left), the one-way agent standard errors tends to be larger, while the one-way shopper standard errors tend to be smaller. As before, the two-way clustered standard errors tend to be closer to the more conservative of the two.

When considering the interaction effect, we find strange results. In particular, IID errors are largest on average, though it is unclear why this is the case. Next largest are the errors clustered one-way on shopper or agent (with ICCs equal, the the larger groups tends to have larger, errors. Finally, two-way clustering provides the smallest of the more consistent errors (i.e., leaving aside one-way market standard errors).

Some observations can also be made from changing the number of shoppers and agents. First, as group size goes up, the variance of the error falls, when clustering one-way on that group and vice-versa (as one would expect). Second, the variance of two-way clustered standard errors mirrors the variance of the *smaller* group.[3] Finally, when considering the IID errors these tend to follow the larger group on average (though they do not change in variance).

2.1.2 Shopper ICC Does Not Equal Agent ICC

Figures 4, 5, and 6 present standard errors when shopper ICC exceeds agent ICC, varying the number of shoppers and agents. Likewise, Figures 7, 8, and 9 present standard errors when agent ICC exceeds shopper ICC, varying the number of shoppers and agents.

Considering the shopper specification when shopper ICC is larger than agent ICC, comparing Figures 2 and 5, we notice an increase in one-way shopper and two-way standard errors. This makes sense since as the ICC for shopper rises from 0.05 to 0.2, we should see a decrease in effective sample size. That is, each observation is less informative. Otherwise, conclusions remain the same as we adjust the number of shoppers and agents. However, the ordering of standard errors tends to be the same. Considering instead when agent ICC is larger (top left of Figure 8) the standard errors are very similar to those in Figure 2. In fact, the one-way standard errors for both shopper and agent are quantitatively similar, as are the IID errors. However, the two-way standard errors increase in size and are smaller (closer to the agent standard errors).

Note that for the factorial specification and unequal ICCs, the results when $\rho_S > \rho_A$ mirror

[3]Thinking about this in terms of Cameron et al. (2011)'s expression of the Standard Errors, this makes sense: $V(\hat{\beta}) = V^S(\hat{\beta}) + V^A(\hat{\beta}) - V^{S\cap A}(\hat{\beta}) = V^S(\hat{\beta}) + V^A(\hat{\beta})$ since there is only one transaction per shopper-agent dyad. If $V^S(\hat{\beta})$ doubles while $V^A(\hat{\beta})$ halves, $V(\hat{\beta})$ will rise when $V^A(\hat{\beta}) \approx V^S(\hat{\beta})$ to start.

those of when $\rho_S < \rho_A$, simply replacing comments about shoppers with agents, and vice-versa. Thus we will consider the case where shopper ICC is large.

Considering the factorial specification, is mirrored for the shopper gender effect (top right). Considering the agent gender effect, when we have the same number of shoppers and agents (Figures 2 and 5) we tend to see similar one-way agent standard errors, but smaller one-way shopper standard errors. Two-way standard errors tend to fall as well. Considering the interaction effect in the factorial specification, the difference in ICCs drives the one-way agent and one-way shopper standard errors apart. While both are still larger than the two-way standard errors on this effect, the cluster with higher ICC has smaller standard errors, all else held equal. In addition, the standard errors for the smaller ICC also have lower variance.

2.2 False Positives

2.2.1 Shopper ICC Equals Agent ICC

In Tables 1, 2, and 3, I present the rate of false positives when shopper ICC equals agent ICC.

Considering the shopper only specification, we see that only the one-way shopper errors are conservative while others are liberal in varying degrees. Of these, agent standard errors do the worst, followed by IID standard errors, market standard errors, and finally two-way standard errors, which are close to balanced, though still consistently liberal. When shoppers are the smaller group, all choices of standard errors tend to be more liberal, and vice versa.

Considering the factorial model, no set of standard errors is conservative for all the estimated effects, which leaves us with tricky decisions. First, when group sizes are equal, clustering on shopper or agent, yields conservative standard errors for terms that are correlated at this level. For example, one-way clustering on shopper is conservative for shopper gender and the interaction of genders. IID is conservative only on the interaction of shopper gender, though this is difficult to account for.[4]

When considering two-way clustering and one-way clustering on market, we see that errors tend to be liberal, though consistently liberal across effects. In general, two-way tend to be lower, though it is surprising that they are not closer to balanced in this case. In the case of one-way on markets, it is clear that this is the result of too few clusters Cameron and Miller (2015). This issue is the result of a bias-variance trade-off. Though market level standard errors should be conservative, they do not converge as quickly as other standard errors and thus produce too many simulations where the market standard error is very small, driving the over-rejection. The

[4]One idea is that because there is only one unique interaction for each dyad in this simulation, this is equivalent to clustering on dyad, which would be the appropriate level of clustering for the correlation observed in this variable. For multiple transactions, we would want to cluster on the dyad (note this is not the same thing as dyadic robust standard errors).

common approach is to utilize wild cluster bootstrap to address this MacKinnon and Webb (2018); Roodman et al. (2019).[5]

Overall, there are limited differences in standard errors by group size. Somewhat unexpectedly, we also see that when we have unequal sized groups, IID errors are conservative for the larger group. Similarly, the performance of one-way standard errors is better overall when they cluster on the smaller group, though they sill overreject (e.g., more than two-way for coefficients on terms related to the group that is not clustered on). The intuition here is not obvious. One would expect the larger group to have less information for each observation, meaning the IID errors would differ more from the robust standard errors.[6]

2.2.2 Shopper ICC Does Not Equal Agent ICC

Figures 8-9 present results when shopper and agent ICCs differ. In general, results change largely in degree as opposed to character (liberal or conservative), though occasionally an effect-standard error will switch. Those standard errors that were conservative for some effect before tend to continue to be conservative for that same effect. However, there are notable changes.

Considering the shopper only specification, when agent ICC increases, the one-way shopper clustering tends to be more conservative. On the contrary, agent standard errors become even worse, particularly when we have few shoppers. On the other hand, when shopper ICC increases, shopper standard errors become more liberal, in these simulations close to 5%, sometimes peeking over. Finally, in this scenario, IID and one-way agent standard errors become considerably worse, reaching their zenith (in a bad way) in the simulation with few shoppers. Here these errors over reject at a rate almost seven times the true size of the test.

Considering the factorial specifications, we have really three scenarios. First, equal sized groups with unequal ICCs; second, larger groups with larger ICCs; and third, smaller groups with larger ICCs. Starting with the first case, equal sized groups. IID standard errors are more conservative for the interaction term and tend to be conservative for groups with more members and smaller ICCs. One-way clustering on agent becomes more conservative for the agent gender term when there is lower ICC within agent (likewise for shopper). Two-way clustering remains very stable, but still is liberal. Second, when larger groups have larger ICCs (see Tables 5 and 9), IID errors tend to be liberal. One-way errors tend to be consistent with the equal groups sizes case. Third, when smaller groups have larger ICCS (see Tables 6 and 8), one-way errors on the

[5]Because of the similarity of the two-way standard errors and one-way market standard errors, a similar consideration might be made for these errors. However, considering the size of the standard errors, it's important to note that these tend to blend the one-way agent and shopper standard errors. This often means they're a blend of a larger and smaller error (for that coefficient) meaning their mean lies in between as does their variance. This would discredit the bias-variance trade-off.

[6]These also rely on low values of the ICC for the term that is conservatively rejected.

smaller group are even more conservative for their like term. Finally, the large group with small ICC rejects conservatively with IID errors. However, the small group with large ICC becomes very liberal, rejecting at more than four times the true size.

References

A. C. Cameron and D. L. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015. ISSN 0022-166X. doi: 10.3368/jhr.50.2.317. URL http://jhr.uwpress.org/content/50/2/317.shorthttp://unionstats.gsu.edu/9220/Camerer-Miller_Clustering_JHR_2015.pdf.

A. C. Cameron, J. B. Gelbach, and D. L. Miller. Robust inference with multiway clustering. *Journal of Business and Economic Statistics*, 29(2):238–249, 2011. ISSN 07350015. doi: 10.1198/jbes.2010.07136.

J. G. MacKinnon and M. D. Webb. The wild bootstrap for few (treated) clusters. *Econometrics Journal*, 21(2):114–135, 2018. ISSN 1368423X. doi: 10.1111/ectj.12107.

D. Roodman, J. G. MacKinnon, M. Nielsen, and M. D. Webb. Fast and wild: Bootstrap inference in Stata using boottest. *Stata Journal*, 19(1):4–60, 2019. ISSN 15368734. doi: 10.1177/1536867X19830877.

A Figures

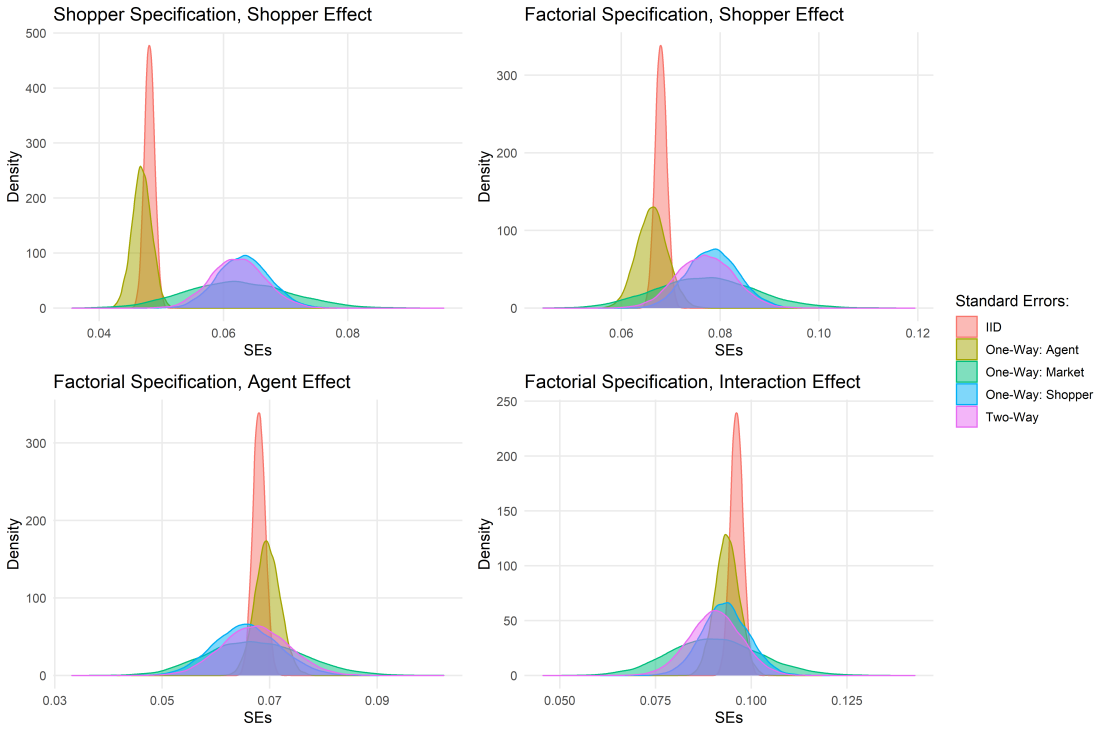

Figure 1: Simulated SEs when $n_S = 4$, $n_A = 16$, $n_M = 30$, $\rho_S = 0.05$, $\rho_A = 0.05$, and $\rho_M = 0$

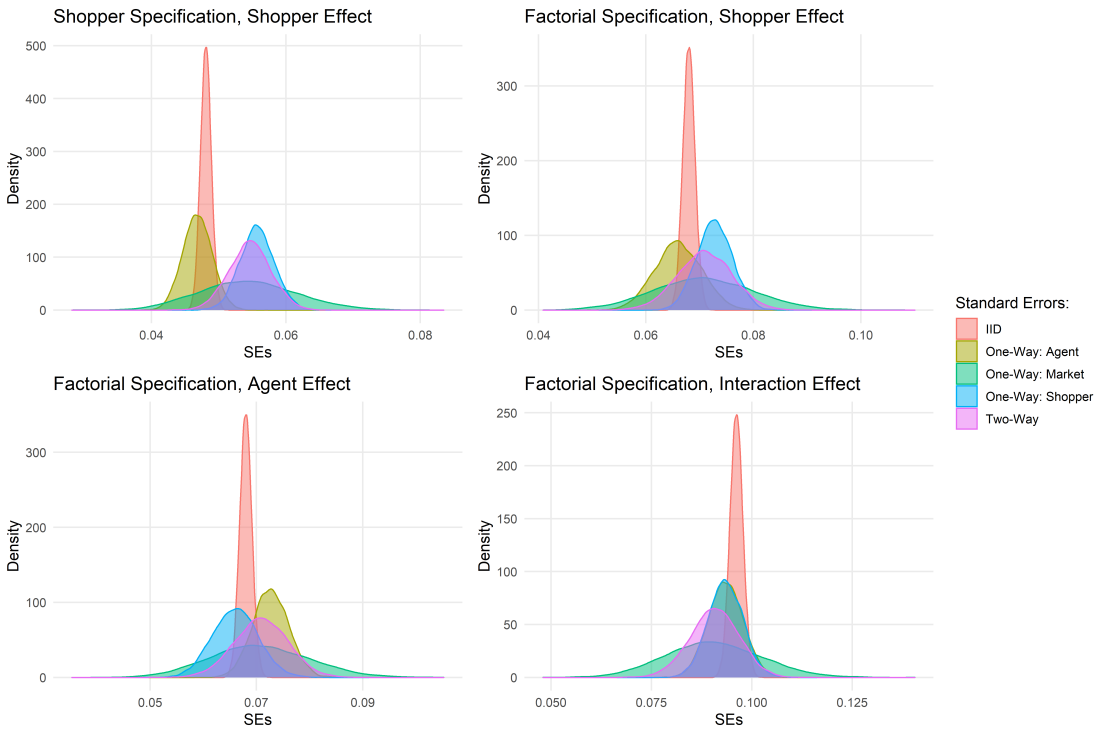

Figure 2: Simulated SEs when $n_S = 8$, $n_A = 8$, $n_M = 30$, $\rho_S = 0.05$, $\rho_A = 0.05$, and $\rho_M = 0$

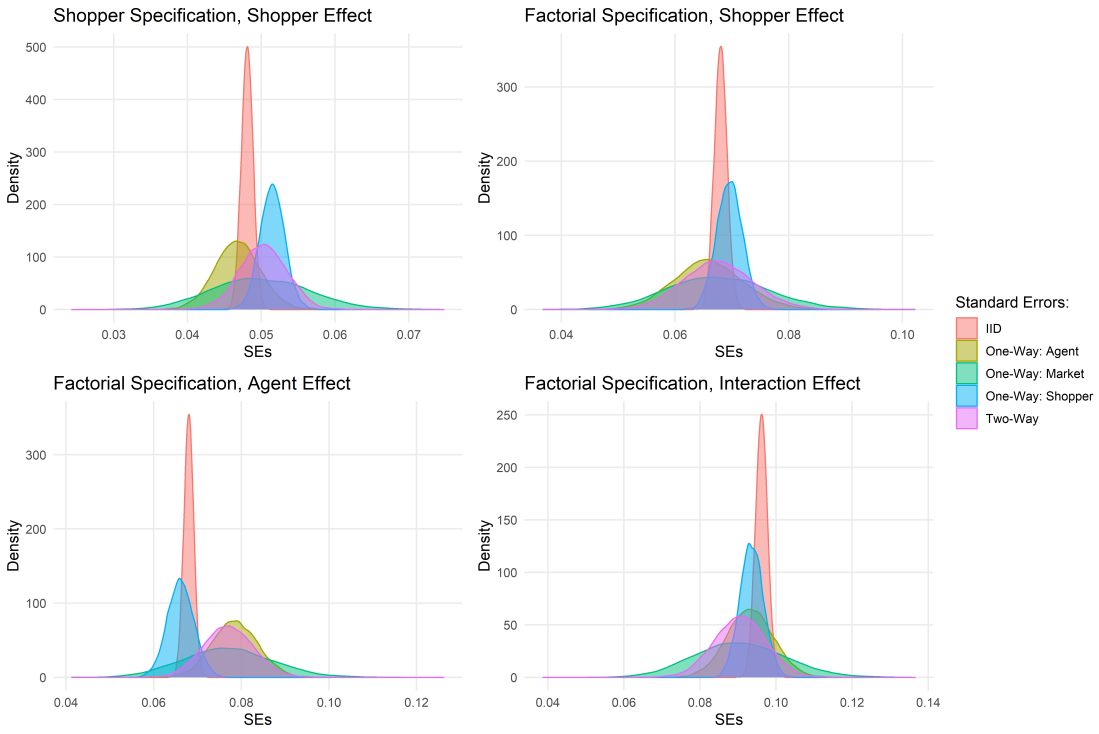

Figure 3: Simulated SEs when $n_S = 16$, $n_A = 4$, $n_M = 30$, $\rho_S = 0.05$, $\rho_A = 0.05$, and $\rho_M = 0$

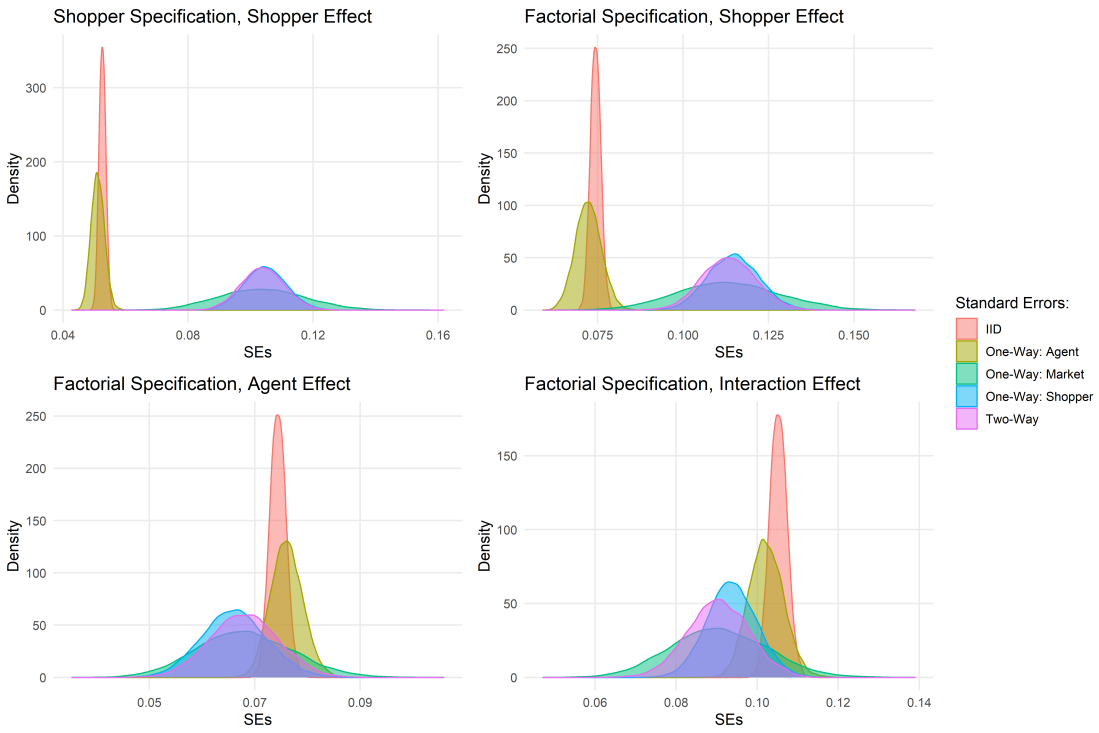

Figure 4: Simulated SEs when $n_S = 4$, $n_A = 16$, $n_M = 30$, $\rho_S = 0.20$, $\rho_A = 0.05$, and $\rho_M = 0$

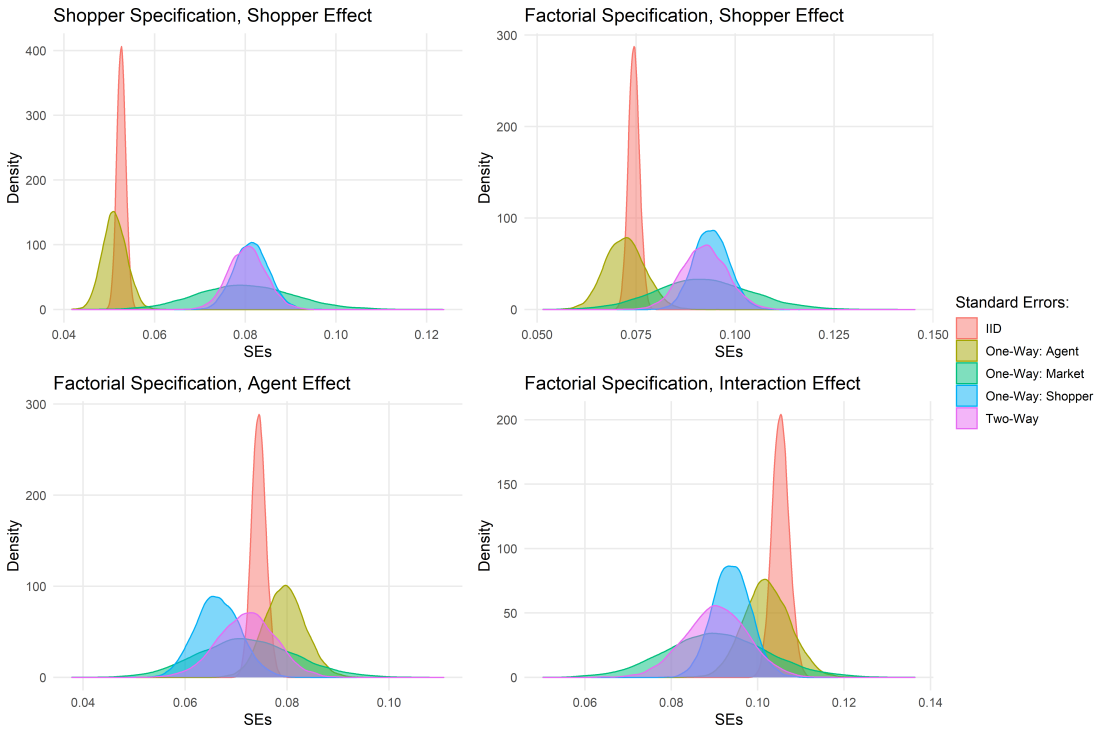

Figure 5: Simulated SEs when $n_S = 8$, $n_A = 8$, $n_M = 30$, $\rho_S = 0.20$, $\rho_A = 0.05$, and $\rho_M = 0$

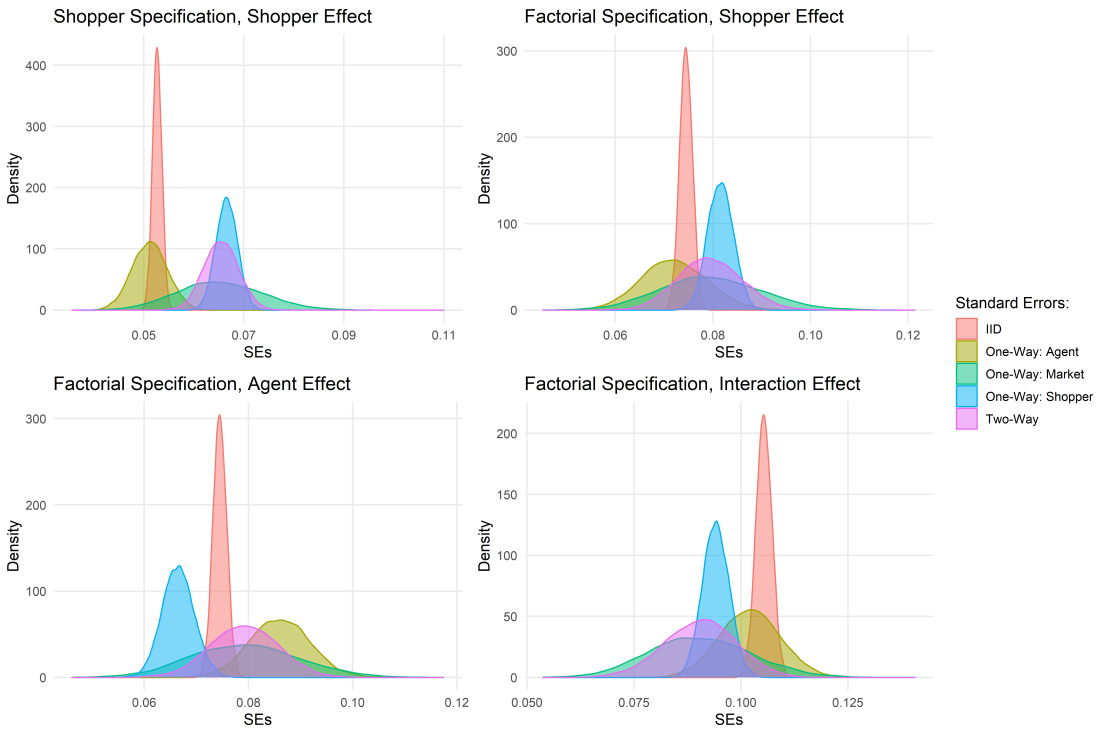

Figure 6: Simulated SEs when $n_S = 16$, $n_A = 4$, $n_M = 30$, $\rho_S = 0.20$, $\rho_A = 0.05$, $\rho_M = 0$

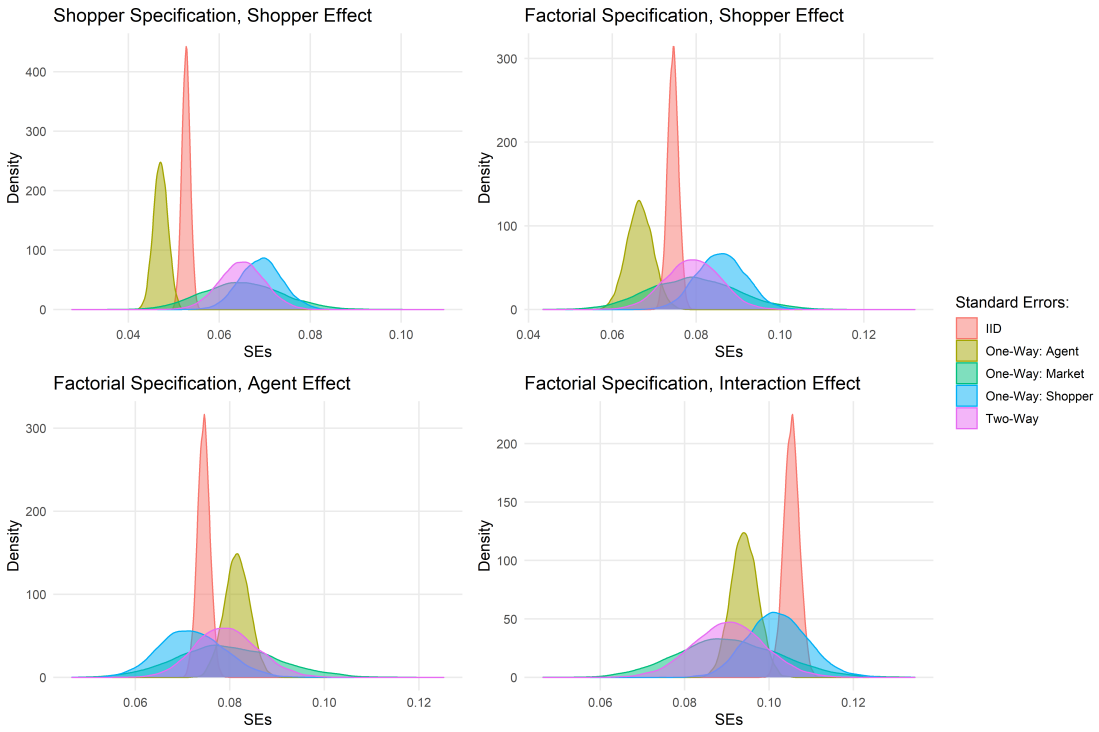

Figure 7: Simulated SEs when $n_S = 4$, $n_A = 16$, $n_M = 30$, $\rho_S = 0.05$, $\rho_A = 0.20$, and $\rho_M = 0$

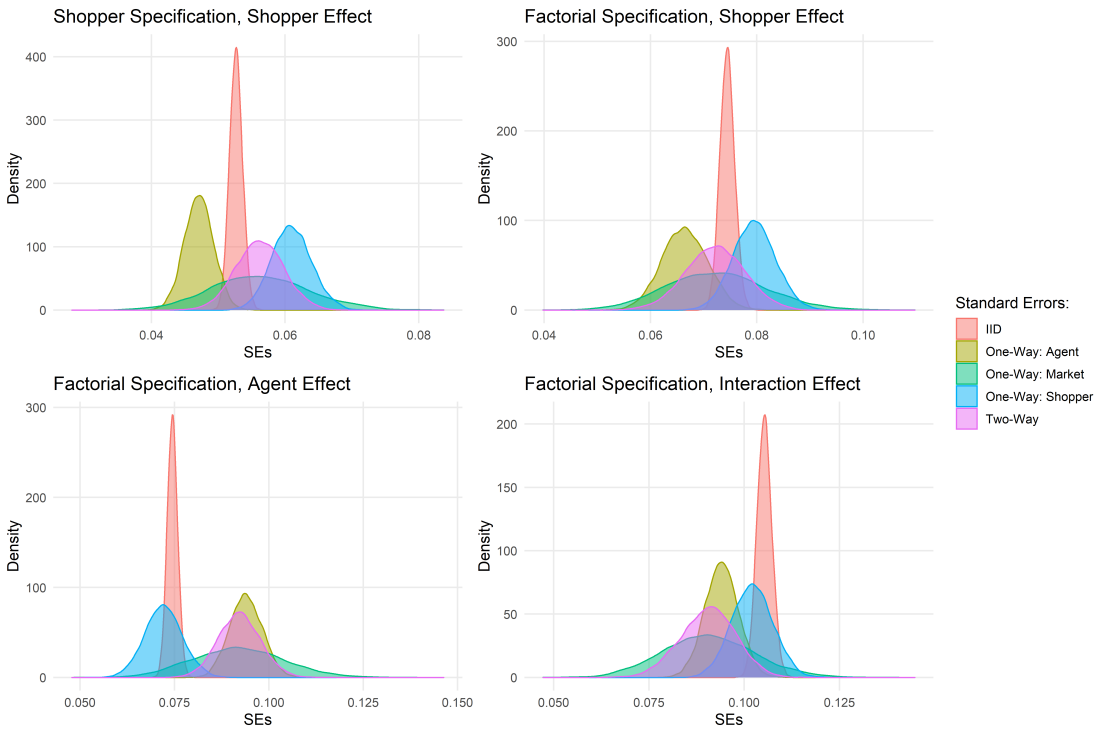

Figure 8: Simulated SEs when $n_S = 8$, $n_A = 8$, $n_M = 30$, $\rho_S = 0.05$, $\rho_A = 0.20$, and $\rho_M = 0$

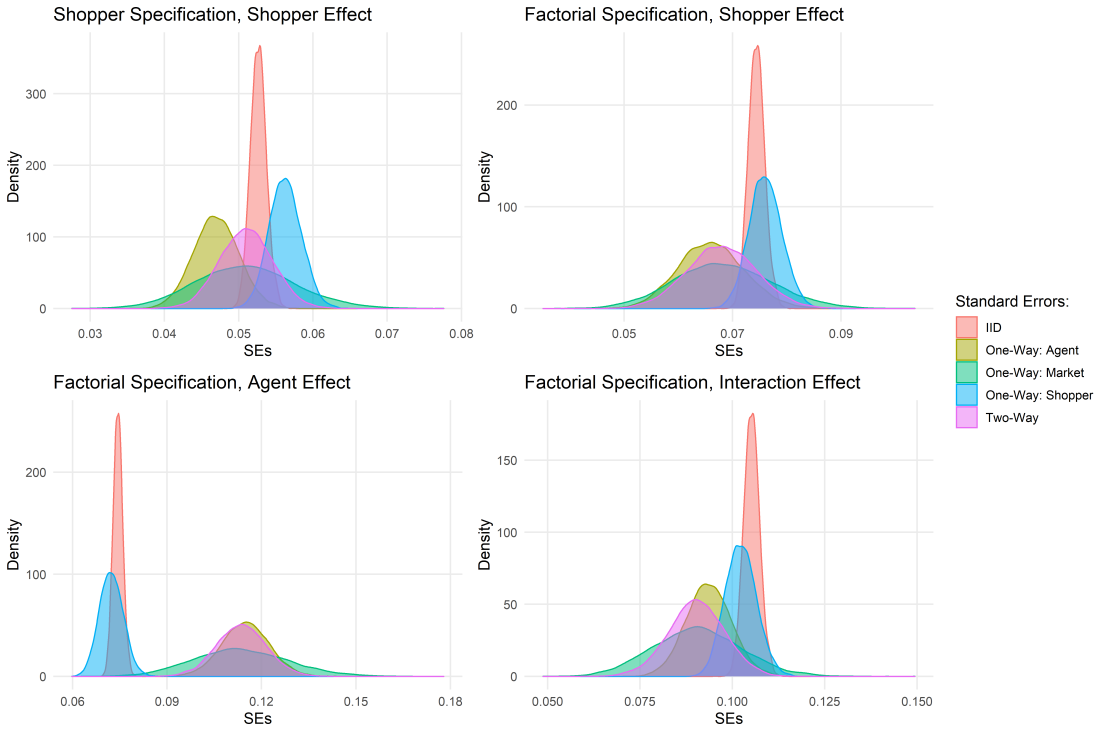

Figure 9: Simulated SEs when $n_S = 16$, $n_A = 4$, $n_M = 30$, $\rho_S = 0.05$, $\rho_A = 0.20$, and $\rho_M = 0$

B Tables

Table 1: False Positives in Audit Studies by Choice of Standard Errors: $n_S = 8$, $n_A = 8$, $n_M = 30$, $\rho_S = 0.05$, $\rho_A = 0.05$, and $\rho_M = 0$.

	Rejection Rate under Null			
	Gender of	Factorial Model: Gender of		
Standard Errors	Shopper	Shopper	Agent	Interaction
IID	8.74	6.04	6.00	**3.78**
One-Way: Agent	9.81	7.02	**4.48**	**4.47**
One-Way: Shopper	**4.84**	**4.57**	7.07	**4.49**
Two-Way	5.40	5.20	5.29	5.33
One-Way: Market	6.13	5.70	5.94	5.80

Table 2: False Positives in Audit Studies by Choice of Standard Errors: $n_S = 4$, $n_A = 16$, $n_M = 30$, $\rho_S = 0.05$, $\rho_A = 0.05$, and $\rho_M = 0$.

	Rejection Rate under Null			
	Gender of	Factorial Model: Gender of		
Standard Errors	Shopper	Shopper	Agent	Interaction
IID	12.50	8.42	**4.97**	**3.85**
One-Way: Agent	13.65	9.20	**4.37**	**4.42**
One-Way: Shopper	**4.63**	**4.80**	6.11	**4.59**
Two-Way	5.03	5.32	5.57	5.37
One-Way: Market	5.57	5.81	5.76	5.89

Table 3: False Positives in Audit Studies by Choice of Standard Errors: $n_S = 16$, $n_A = 4$, $n_M = 30$, $\rho_S = 0.05$, $\rho_A = 0.05$, and $\rho_M = 0$.

	Rejection Rate under Null			
	Gender of	Factorial Model: Gender of		
Standard Errors	Shopper	Shopper	Agent	Interaction
IID	5.93	**4.53**	8.32	**3.67**
One-Way: Agent	6.77	5.84	**4.72**	**4.38**
One-Way: Shopper	**4.41**	**4.05**	9.48	**4.36**
Two-Way	5.02	5.20	5.38	5.04
One-Way: Market	5.76	5.33	5.76	5.70

Table 4: False Positives in Audit Studies by Choice of Standard Errors: $n_S = 8$, $n_A = 8$, $n_M = 30$, $\rho_S = 0.05$, $\rho_A = 0.2$, and $\rho_M = 0$.

	Rejection Rate under Null			
	Gender of	Factorial Model: Gender of		
Standard Errors	Shopper	Shopper	Agent	Interaction
IID	6.64	**4.34**	11.5	**2.54**
One-Way: Agent	9.98	7.33	**4.63**	**4.60**
One-Way: Shopper	**3.56**	**3.14**	13.07	**3.12**
Two-Way	5.20	5.37	5.13	5.74
One-Way: Market	5.74	5.80	5.94	6.08

Table 5: False Positives in Audit Studies by Choice of Standard Errors: $n_S = 4$, $n_A = 16$, $n_M = 30$, $\rho_S = 0.05$, $\rho_A = 0.2$, and $\rho_M = 0$.

	Rejection Rate under Null			
	Gender of	Factorial Model: Gender of		
Standard Errors	Shopper	Shopper	Agent	Interaction
IID	11.92	6.70	6.76	**2.22**
One-Way: Agent	16.20	10.29	**4.61**	**4.44**
One-Way: Shopper	**4.09**	**3.66**	8.12	**2.89**
Two-Way	5.55	5.68	5.43	5.56
One-Way: Market	6.10	5.85	5.86	5.66

Table 6: False Positives in Audit Studies by Choice of Standard Errors: $n_S = 16$, $n_A = 4$, $n_M = 30$, $\rho_S = 0.05$, $\rho_A = 0.2$, and $\rho_M = 0$.

	Rejection Rate under Null			
	Gender of	Factorial Model: Gender of		
Standard Errors	Shopper	Shopper	Agent	Interaction
IID	4.57	3.25	19.79	**2.15**
One-Way: Agent	7.40	6.03	**4.77**	**4.32**
One-Way: Shopper	**3.40**	**2.87**	21.33	**2.73**
Two-Way	5.51	5.36	**4.98**	5.22
One-Way: Market	6.00	5.73	5.59	5.36

Table 7: False Positives in Audit Studies by Choice of Standard Errors: $n_S = 8$, $n_A = 8$, $n_M = 30$, $\rho_S = 0.2$, $\rho_A = 0.05$, and $\rho_M = 0$.

	Rejection Rate under Null			
	Gender of	Factorial Model: Gender of		
Standard Errors	Shopper	Shopper	Agent	Interaction
IID	20.74	11.96	**4.54**	**2.51**
One-Way: Agent	22.18	13.09	**3.33**	**3.11**
One-Way: Shopper	5.12	**4.89**	7.85	**4.59**
Two-Way	5.42	5.39	5.54	5.69
One-Way: Market	6.15	6.13	6.01	6.17

Table 8: False Positives in Audit Studies by Choice of Standard Errors: $n_S = 4$, $n_A = 16$, $n_M = 30$, $\rho_S = 0.2$, $\rho_A = 0.05$, and $\rho_M = 0$.

	Rejection Rate under Null			
	Gender of	Factorial Model: Gender of		
Standard Errors	Shopper	Shopper	Agent	Interaction
IID	32.49	20.26	**3.33**	**2.27**
One-Way: Agent	33.76	21.65	**2.97**	**2.83**
One-Way: Shopper	**4.91**	**4.73**	6.08	**4.59**
Two-Way	5.01	5.06	5.29	5.6
One-Way: Market	5.45	5.74	5.54	5.79

Table 9: False Positives in Audit Studies by Choice of Standard Errors: $n_S = 16$, $n_A = 4$, $n_M = 30$, $\rho_S = 0.2$, $\rho_A = 0.05$, and $\rho_M = 0$.

	Rejection Rate under Null			
	Gender of	Factorial Model: Gender of		
Standard Errors	Shopper	Shopper	Agent	Interaction
IID	11.49	7.07	6.75	**2.72**
One-Way: Agent	12.89	8.5	**3.61**	**3.28**
One-Way: Shopper	**4.39**	**4.64**	10.68	**4.79**
Two-Way	**4.69**	5.61	5.44	6.08
One-Way: Market	5.65	5.96	6.02	6.34