To Match or Not to Match? Statistical and Substantive Considerations in Audit Design and Analysis*

April 4, 2017

Mike Vuolo
The Ohio State University

Christopher Uggen
University of Minnesota

Sarah Lageson
Rutgers University

* Corresponding author: Mike Vuolo, Department of Sociology, The Ohio State University, 238 Townshend Hall, 1885 Neil Avenue Mall, Columbus, OH 43210; email: vuolo.2@osu.edu.

To Match or Not to Match? Statistical and Substantive Considerations in Audit Design and Analysis

In audits, as in all experiments, researchers are confronted with choices about whether to collect and analyze repeated measures on the unit of analysis. In typical social science practice, this decision often involves consideration of whether to send single or multiple auditors to test for discrimination at a site that represents the unit of analysis, such as employers, landlords, or schools. In this chapter, we provide tools for researchers considering the statistical and substantive implications of this decision. For the former, we show how sample size and statistical efficiency questions hinge in large part on the expected concordance of outcomes when testers are sent to the same unit or site. For the latter, we encourage researchers to think carefully about what is gained and lost via matched and non-matched designs, particularly regarding the finite nature of certain populations, resource constraints, and the likelihood of detection in the field. For both approaches, we make recommendations for the appropriate statistical analysis in light of the given design and direct readers to software and code that may be helpful in informing design choices.

To Match or Not to Match? Statistical and Substantive Considerations in Audit Design and Analysis

In prior work on choosing a sample size for a paired audit (Vuolo, Uggen, and Lageson 2016), we recommended that researchers consider unmatched designs under certain research conditions. This would typically involve sending one (unmatched) randomly assigned treatment/control "tester" to a single experimental unit (e.g. employer, school, landlord), rather than sending multiple testers to the same unit. In this chapter, we consider this recommendation in more detail, specifically from the perspective of sample size requirements and statistical efficiency. We weigh the advantages and disadvantages of matched and unmatched approaches statistically and substantively. First, we compare sample size requirements for matched and unmatched designs. Then, we examine empirical data from our own audit (see Uggen et al. 2014) and a series of simulations to contrast the statistical efficiency of modeling approaches for matched and unmatched audits, including predictors that vary both within and between experimental units. Next, we repeat this exercise for a hypothetical case with a much different distribution of outcomes than our empirical audit results. The lesson from these exercises is that the more efficient approach rests heavily on the expected degree of concordance of outcomes. Then, we discuss substantive considerations for matching that should be taken into account alongside efficiency. Finally, in light of these results, we offer further recommendations for the specific choice of whether to match or not.

1. Sample size requirements for matched and unmatched approaches

Issues of statistical power are central in determining the proper sample size to detect effects, and thus are crucial to careful research design that maximizes the chance for meaningful

results and well-expended resources. Although there are several values that determine power, one intuitive way to understand power is in terms of the magnitude difference. That is, at what magnitude difference in the population does a particular sample size have a reasonable chance (typically 80 or 90 percent) that a statistically significant effect ($p < .05$) will be detected in a given sample? In determining an appropriate sample size then, we select a magnitude difference and determine the minimum sample size that would result in a significantly significant effect ($p < .05$) in 80 to 90 percent of samples, should a difference of that magnitude or greater exist in the population. In this chapter, we examine this question for audit studies that employ either matched or unmatched designs. Table 1 shows audits from our disciplines of sociology and criminal justice. While this list is by no means exhaustive, it does show a rapid increase in the number of published audit studies since Pager's (2003) landmark research. Importantly here, while 9 of the 10 audits prior to 2015 employed matched designs, there is a recent rise in unmatched designs (4 of 11 since 2015). In light of this current landscape of audits, this is a particularly opportune moment to assess the differences in these two design approaches.

[Table 1 here.]

Table 2 provides notation for the cases of matched and unmatched experimental designs. Throughout this chapter, we use the example of a 2×2 experiment, where there is a single treatment and control condition both sent to an experimental unit where a response is measured either affirmatively or negatively. In our empirical matched example (Uggen et al. 2014), we sent two job applicants (testers) to employers (experimental unit) and measured whether they were called back by the employer (affirmative response) or not called back by the employer (negative/no response) when one presented an arrest record (treatment) and the other presented a clean record (control). In the table, p_{11} represents the proportion of experimental units where

both testers received an affirmative response, while p_{00} is the proportion where both testers received a negative (or no) response. Together, $p_{11} + p_{00} = p_{CC}$, or the total concordance. By contrast, p_{10} represents the proportion of units where the control received an affirmative response and the treatment did not (that is, the cell where discrimination is observed), while p_{01} is the proportion where the treatment received an affirmative response when the control did not. Together, $p_{10} + p_{01} = p_{DD}$, or the total discordance. The difference in the discordance is the object of the test of statistical significance for matched designs and is known as McNemar's test (McNemar 1947). In prior work (Vuolo et al. 2016), we demonstrated that the sample size requirements for a matched audit of a given power level and alpha are based not only on the difference in the discordant proportions ($p_{10} - p_{01}$), but also the total amount of concordance (p_{CC}). We emphasized that the distribution of total concordance p_{CC} across the two constituent concordant cells was irrelevant for sample size calculations; only the total mattered.

[Table 2 here.]

In an unmatched design, only the marginal distribution, or the proportions in the "Total" column and row, is observed. But the difference in the marginals will always equal the difference in the two discordant cells, or $p_{10} - p_{01} = p_{1+} - p_{+1}$ (see Vuolo et al. 2016 for a proof). So regardless of whether matched or not, the observed percentage point difference is the same. The sample size requirements to detect that effect, however, may vary greatly. In fact, while the total concordance is the primary consideration for matched designs, the required sample size for an unmatched design must take the values of the two concordant cells into account because as they change, the marginal values p_{1+} and p_{+1} change, despite the discordant cells and $p_{10} - p_{01} = p_{1+} - p_{+1}$ remaining the same.

In our prior recommendation regarding matching versus non-matching (Vuolo et al. 2016:295), we stated that smaller sample sizes are required for matched tests when there is greater concordance, while lower sample sizes are required for independent tests when there is greater discordance. This result is straightforward: when the experimental unit (say, the employer in a study of job discrimination) has little effect on the results of the test, it is irrelevant whether one sends testers to the same unit. We based this conclusion on results from Donner and Li (1990), who showed that the size requirements for the unmatched Pearson's chi-square test are related to a matched test via a weight that measures intraclass correlation. Based solely on sample size, the key to deciding between a matched and unmatched approach depends on whether this weight exceeds or falls below 1.

As expressed in the formula (see Appendix B in Vuolo et al. 2016), this is equivalent to asking whether concordance is above or below .5. This result is intuitive, as .5 marks the threshold at which the effect of the experimental unit (e.g. an employer) becomes more or less influential. That is, it represents the point at which half of the tests had either the same or different outcome occur at the same experimental unit. When more than half of the tests have the same outcome, the experimental unit exerts an effect and matching is preferred, and vice versa. Mathematically, this weight produces lower sample size values for the matched case when concordance is above .5 and the matched design is preferred. The opposite is true when the concordance falls below .5 because the weight would then produce lower sample size values for the unmatched design, which would then be preferable.

Our prior work focused primarily on matched designs. We here build upon that study by comparing such matched designs against unmatched designs. We begin by further examining sample size requirements, which results in some caveats to our prior recommendation. According

to the weight described above, a 50-50 concordance-discordance split designates the point at which one design is favored over the other. While we find that the unmatched case is always preferable in terms of sample size when concordance is below .5, there are still instances where the unmatched case requires lower sample size even when the concordance is above .5. As noted above, the sample size required for McNemar's test for matched designs for a given p_{10} and p_{01} does not depend on the breakdown of the two concordant cells, as only the total p_{CC} matters. For the unmatched case, this breakdown does matter and produces different sample size requirements depending on the marginal distribution, with the well-known result that a higher sample size is required as the distribution approaches values of 50-50, or a random coin flip. Depending on how far the marginal distribution departs from this even split, the sample size requirements for the unmatched case are lower than for the matched case. In our prior work, we emphasized that only total concordance mattered for sample size calculations in matched designs. The distribution across the concordant cells, however, becomes an important factor when deciding between matched and unmatched designs.

We show this result graphically in Figure 1, with each panel displaying sample size requirements for a power of .9 (equivalently a Type II error of $\beta = .1$), Type I error of $\alpha = .05$, and population difference of 5 percentage points between affirmative responses for the treatment and control. Substantively then, the figures display the sample size required to have a 90% chance of observing a sample that would result in a statistically significant effect at the .05 level for a population difference of a given amount (in this example, 5 percentage points). To compute sample sizes, we use the formula (Rosner 2011:384-86) and function for the statistical software program R (R Core Team 2015) presented in our prior work (Vuolo et al. 2016). For the unmatched case, we use the R function "power.prop.test" (for derivation and more on this R

function, see Chen and Peace 2011:163-66). We note that our McNemar's function calculates the number of experimental units, and the unmatched sample size function calculates the number of units per group. That is, the total observations for both are actually twice the amount shown. We return to substantive implications of this below.

Panel A shows sample size requirements for a distribution where $p_{01} = .05$, $p_{10} = .10$, and $p_{CC} = .85$. In other words, the control receives an affirmative response at 10% of experimental units when the treatment did not, and the treatment receives an affirmative response at 5% of experimental units when the control did not. The remaining 85% of experimental units are concordant, which means both testers uniformly received an affirmative or negative response. For McNemar's test for matched pairs, the sample size requirement is the same regardless of the values in the two concordant cells, represented by the horizontal line with a value of 603.

The values of the concordant cells do, however, matter for the unmatched test of proportions, which is shown by the curve. The x-axis is the marginal value p_{1+}, or the total proportion of affirmative responses for the control regardless of the outcome of the treatment test (again, this is all that would be observed in the unmatched case). Since the gap in the marginals is the same as the gap in the discordant cells, the x-axis could also represent p_{+1} (the total proportion of affirmative responses for the treatment regardless of the outcome of the control test) by simply subtracting .05. The implications for sample size are clear and corroborate the results of the weight discussed above: the matched case virtually always requires a lower sample size with a concordance of 85%.

[Figure 1 here.]

Perhaps unexpectedly, as we move across the panels and the concordance decreases, however, there are scenarios in which the unmatched is preferred from a sample size perspective

even when the concordance is below .5. In Panel B with 75% concordance, the required sample size for McNemar's test represented by the horizontal line is 1035, but at the ends of the marginal distribution, a small proportion of possible distributions fall below the horizontal line, thus favoring the unmatched design. As we move across concordance totals through the panels, more of the unmatched curve falls below the required sample size for the matched design. Once the total crosses .5, the curve is completely below the required sample size for the matched case. In Panel E with 55% concordance, the maximum of the unmatched curve, with a value of 2100, is completely below the required sample size of 2305 for the matched case. The curve actually remains constant across the panels, but there is less of the curve to display because the possible marginal distributions decrease with concordance. We illustrate these by showing each of the panels together in Panel F.

Figure 2 changes the percentage point difference for the discordant cells and the marginals to .15. A virtually identical pattern is observed, but of course with a y-axis that includes much lower sample sizes given the larger discordant difference and one fewer panel since the concordance starts at a lower value of .75. Again, at the highest concordance, the matched case always requires lower sample size (Panel A). After crossing .5 concordance, the unmatched yields lower sample size requirements (Panel D), but there are scenarios in between where the unmatched case still requires lower sample size (Panel B and C).

[Figure 2 here.]

2. Statistical efficiency in matched and unmatched designs

A more statistically efficient design or experiment is one that requires fewer observations. Paired designs are often encouraged due to a widely-held belief, sometimes made

explicit in introductory texts (e.g., Kramer 1991; Dalgaard 2008; Shih and Aisner 2016), that they are more statistically efficient than unmatched designs. This perception is likely based on the case of a paired *t*-test, where paired designs with a continuous outcome are often stated as more statistically efficient, even though this is not always the case (Hedberg and Ayers 2015). With McNemar's test as the analogue to a paired *t*-test when the outcome is dichotomous, clearly from what we showed above, the same claim about the efficiency of this design is also misplaced. More specifically, the matched case is only more efficient when concordance is higher than .5, and still only with certain marginal distributions.

In practice, an estimator is more efficient if its standard error, or the standard deviation of its sampling distribution, is smaller. Thus, in this section, we consider the standard errors of the coefficients from standard modeling approaches that would follow from matched and unmatched designs. For unmatched designs, the model is simply a standard logistic regression. For the matched case, we display both a logistic regression with cluster-corrected standard errors and a multilevel logistic regression of observation nested within experimental unit, as both are often seen in audit studies. As would be expected from the sample size figures above, the standard errors are smaller in the matched case when concordance is high (> .5) and smaller in the unmatched case when concordance is low (< .5). Beyond demonstrating this result, however, we also consider two related components that are often important in an audit that are not reflected in the preceding section: (1) a randomized blocking variable at the experimental unit level (as opposed to an observed covariate at the experimental unit level) that thus only exhibits between-unit effects; and (2) the interaction of that variable with the treatment condition. Thus, while the calculations in the prior section solely consider the efficiency of the focal treatment (that which would vary within-units in the matched case), here we consider both blocking effects and the

(cross-level in the matched case) interaction. As is shown below, we find the efficiency of the focal treatment and interaction to be at odds with that of the blocking effect.

For example, in our own audit (Uggen et al. 2014), half of the experiments were White tester pairs and half were African American tester pairs. In the interest of testing the arrest record, these pairs were always sent to the same employer, such that the effect of race was actually at the level of the experimental "block" (here, the employer level). This was also the design of Pager's (2003) Milwaukee audit (in which same-race pairs were sent to employers, with one of the testers presenting a felony conviction) and Correll, Benard, and Paik's (2007) audit (in which same-gender pairs were sent to employers, with one tester signaling parenthood status). In our prior work on sample size for matched audits (Vuolo et al. 2016), we stated that the blocks should be considered separate experiments for the sake of sample size calculations (that is, for example, one experiment for Whites and one for African Americans). Since we are often interested in the effect of this blocking variable, as well as its interaction with the treatment, we now consider this in more detail.

We begin by considering the efficiency of a case with more concordance by using the empirical results from our own audit (Uggen et al. 2014). As described above, we sent same-race pairs to employers to test the effect of an arrest record on employer callbacks. The top of Table 3 shows the results of the audit for the whole sample. From the discordant cells, 13% of employers called back the control with the clean record but not the treatment with the arrest record, while 9% of employers actually called back the tester with the record but did not call back the tester without a criminal record. This amounts to a difference of $p_{10} - p_{01} = .04$. The concordance is high at $p_{CC} = p_{11} + p_{00} = .20 + .58 = .78$. Thus, 78% of employers provided the same response to both testers who applied for jobs at their workplace, either calling both or neither job

applicants, regardless of criminal record. These specific values of p_{11} and p_{00} produce marginal values of $p_{1+} = .33$ and $p_{1+} = .29$, which represent the only values that would be observed in the unmatched case. The frequencies in the table represent employers, such that there are 300 total employers and 600 observations. The distribution of callbacks for the African American and White testers are also shown in the table.

In the following exercise, we assume that our original audit data collection constitutes the population of employers and draw samples from within to demonstrate efficiency. We know the population parameters of the statistical models for the matched case (as that was our original design), but not for the unmatched case. We therefore simulate an unmatched design by drawing just a single observation from each employer, while maintaining the balance of race and arrest record. The result is 150 observations per treatment/control group (300 total observations), with equal numbers per race. This results in 75 of each race-arrest combination. To compare matched and unmatched designs of similar sample sizes, we also simulate the matched case by randomly selecting 150 pairs among the 300, while again keeping race balanced.

We repeated this exercise 100,000 times to approximate the sampling distribution. Thus, for each simulation, the value of the coefficients is recorded as an observation in the sampling distribution. The standard deviation of these 100,000 values provides the standard error. In what follows, we focus primarily on the comparative sizes of the standard errors, and not the coefficients or their statistical significance. We restrict our discussion to the standard errors (see Uggen et al. 2014 for the interpretation of the coefficients in our original audit).

Table 4 displays the results of these simulations. The lowest standard error for a given model is indicated by bold type. Three models are shown in the rows: one with the effect of the treatment, one that adds an employer-level randomized variable, and the interaction of both. We

begin by comparing the unmatched logit model (shown in the first column) to the matched logit model with cluster-corrected standard errors (shown in the second and third columns). Model 1 confirms our sample size calculations for cases of high concordance: the matched sample has a lower standard error according to the cluster-corrected model than the unmatched sample. Model 2 reveals an interesting finding: the employer-level effect for White testers actually has a lower standard error for the unmatched case. In the case of high concordance then, the cluster-level effect is more efficient in the unmatched case than in the matched case.

Figure 3 depicts the sampling distributions for the simulated Model 2's, with histograms graphing the frequency of the various simulated coefficient values across 100,000 draws of 300 observations. The lower standard error is reflected in the tighter distributions for White in the unmatched case, and for the arrest record in the matched case. Turning to the matched hierarchical model, the coefficients and standard error are larger. These differences likely emerge because, with a binary outcome, the hierarchical model represents the unit-specific model, whereas the clustered standard errors represent the population averaged model. While technically less efficient, the coefficients and standard error for the hierarchical model are proportional to the cluster correction, such that the same inferential conclusions (i.e., *p*-values) are reached.

[Table 4 and Figure 3 here.]

Although randomization results in no covariation between the treatment and the cluster-level effect, there is still the potential for a significant interaction between these two measures. For an unmatched design, this interaction reflects four unique treatment/control categories, as each combination is randomly assigned to a single employer. For the matched case, this cross-level interaction represents the difference-in-difference between the treatment and control at employers that were randomly assigned a race pair. Regarding efficiency for Model 3 in Table 4,

we find that the interaction is more efficient for the matched design when concordance is high, which also remains the case for the main effect of the record treatment. The standard errors for the main effect of the employer-level race effect are virtually identical regardless of matching. We similarly show the sampling distribution for Model 3 in Figure 4. As expected, there are tighter distributions in the matched case for the arrest record treatment and the interaction, but nearly identical distributions for the main effect of White.

[Figure 4 here.]

We next consider the case of low concordance. As demonstrated above, the unmatched case is more efficient according to sample size calculations. This result is intuitive: for matching to matter, the experimental unit (e.g. employers) must respond at least somewhat similarly to the matched pairs (e.g., testers). Lacking such an empirical dataset, we created a mock dataset with low concordance. However, the example of employers as the experimental unit is inadequate in this case, as it would imply that at least 50% of employers called back one tester, but not the other (and would also include cases where the employer called back both). Such callback rates would be unrealistically high, based on every published audit of employers of which we are aware (see Vuolo et al. 2016 for a summary).

By contrast, audits of *landlords* typically have much higher callback rates than employer audits and provide a published example of an unmatched design (Lauster and Easterbrook 2011). Given this, our mock example assumes we have two apartment-seekers (testers) who ask to see an advertised housing unit from a landlord (experimental unit). The outcome of interest is whether or not they were called back by the landlord to tour the unit (considered an affirmative response). The experimental manipulation is the presence of children, signaled by one tester stating they were interested in the unit for their family (treatment) and the other tester not

mentioning any family (control). To the extent that landlords view children as a risk (Desmond 2016), we would expect some level of discrimination. Children are hardly disqualifying, however, as there are many likely instances in which a landlord might prefer a family to a single individual. Finally, in contrast to our empirical example of job hunting, we expect landlords to be more eager to show units to prospective tenants, relative to employers' tendencies to call back prospective employees. As before, race is used as a cluster-level effect in the matched case, meaning that same-race pairs inquire about a single unit.

While there are many distributions among the four cells that we could have chosen, for the sake of an example, we chose the population distribution in Table 5 with a concordance of $p_{CC} = .25$. From the top of the table depicting the whole sample, 45% of those who did not mention family were invited to tour the housing unit when the family was not, while 30% of those who mentioned a family were offered a tour when those who presented as single were not. This amounts to a difference of $p_{10} - p_{01} = .15$. For the concordance, we assumed that landlords called back neither inquiry for a tour (p_{00}) 18.7% of the time, and both inquiries (p_{11}) 6.3% of the time. These specific values of p_{11} and p_{00} produce marginal values of $p_{1+} = .513$ and $p_{1+} = .363$, which again represent the only values that would be observed in the unmatched case. The lower two panels of Table 5 distribute these case over the landlord-level effect of race, producing values that imply racial discrimination, but with low concordance. We recognize that this hypothetical example represents a case where there are strong preferences for either the treatment or control (as might be the case, for example, in retirement communities or college campuses, in which housing is segregated by family status), but emphasize that this is only an illustrative case. To offer another hypothetical example, a similarly divided response could occur if testers signaled their partisan political affiliation when applying for jobs. Just as children may

be favored or disfavored by landlords, some employers would have a preference for Republicans and others would discriminate against them.

[Table 5 here.]

Table 6 presents simulated models analogous to those presented above. Looking across the columns to identify the lowest standard errors (again shown in bold), the results clearly indicate that the most efficient estimators in the low concordance case are opposite of those observed in the high concordance case. In this case, the two matched modeling approaches yield almost identical results and lead to the same conclusions. In this simulation, the hierarchical and clustered standard error approaches are much more similar, likely because the unit-specific and population average models converge when there is little effect of the unit (as with low concordance). With low concordance, the lowest standard errors for the family treatment effect and the interaction belong to the unmatched design. Clearly, these decisions have implications for the ability to detect a significant effect, precisely the point of a priori power analyses. Any matching efficiency benefit for the main treatment of interest (i.e., the treatment condition researchers would vary within an experimental unit) disappears when the concordance is low. As further evidence of the importance of concordance, the White race effect in Model 2 of Table 6 has the lowest standard error when a matched design is used (whereas it was lowest for the unmatched design in Model 2 of Table 4). Figure 5 and 6 reiterate this efficiency for the matched design by again displaying the simulated sampling distribution histograms.

[Figure 5 and 6 here.]

3. Substantive considerations to matching

Until this point, we have considered the benefits of matched and unmatched approaches in purely statistical terms, specifically power and efficiency. In this section, we discuss substantive considerations that may lead researchers to prefer one approach over the other. That is, statistical considerations must be weighed within the context of conducting a real-world social experiment in which non-statistical contingencies typically arise.

The first consideration stems from the number of experimental units to be sampled relative to the total number of observations. In the matched design, two testers are sent to n experimental units, such that the number of observations is $2n_m$ but the total experimental units to be sampled remains n_m (m for matched). In the unmatched design, n_u represents the number of observations per group, such that both the number of observations and the total experimental units to be sampled is $2n_u$ (u for unmatched). This result does not imply that matched designs take twice as many experimental units as an unmatched design because, as shown in detail above, they do not require the same sample size n to detect the same difference in the population. That is, $n_m \neq n_u$, except where the horizontal line crosses the curve in Figures 1 and 2. Rather, researchers should calculate the required n for each design, keeping in mind that, if the unmatched design is more efficient with lower n, they will need to sample $2n_u$ units.

The efficiency comparison is calculated on the difference between n_m and n_u, but the need to consider the availability of $2n_u$ units to sample is a substantive consideration, not a statistical one. Whether this is of concern depends, in part, on the total population of experimental units. In correspondence studies of employers that send only a fictional application with no live tester and use several geographic locations (e.g., Wright et al. 2013; Tilcsik 2011; Bertrand and Mullainathan 2004), the population and subsequent sample is typically quite large, such that this is rarely a concern. In a single labor market, however, the population may be more

limited. Moving away from employers, one can imagine a case where the required $2n_u$ could exceed the elements of the population. For example, in an audit of admission to medical schools, there are only so many elements of the population. Thus, despite the efficiency gains in the unmatched case, the matched case still might be preferred because it requires fewer experimental units to be drawn from this limited population.

Another limiting factor concerns resources, such as when the audit uses live testers (e.g., Pager 2003 on job applicants) or when the audits cost money (e.g., Stewart and Uggen 2016 on college admissions). In such a case, the researcher typically only has a budget delineated for a predetermined number, or is seeking funds for a given number. With live testers or costly audits, the total number of observations matters greatly because the same funds will be spent to send testers to $2n$ tests regardless of whether they are sent to n_m and n_u employers. Given resource constraints, the number of proposed total observations is often more restrictive due to the prohibitive cost of compensating live testers or paying for applications. Thus, the likelihood of exceeding the elements of the population when sampling is likely low. Here, we would recommend not using the substantive consideration of $2n_u$ compared to n_m. Rather, given the overall lower sample size when resources are limited, the ability to detect a statistically significant effect should take precedence, such that the efficiency comparison of n_m and n_u should be more important.

An additional substantive consideration concerns the possibility of being discovered or "caught" when conducting an audit that relies on deception. Were such an audit to become known to the experimental units (see, e.g., Gaddis's (2015) discussion of educational credentials in his pilot) or the public (e.g., if a college admissions audit was showcased in the *Chronicle of Higher Education* when researchers were still in the field), results could be biased or

contaminated by this information. Moreover, researchers are often interested in testing more than two treatment levels, as is often the case with race and ethnicity (e.g. Pager, Western, and Bonikowski 2009; Gaddis and Ghoshel 2015b), or multiple treatments, such as criminal record and race (Pager 2003; Uggen et al. 2014), race and skill (Bertrand and Mullainathan 2004), and gender and parenthood (Correll et al. 2007) in employer audits. There are two strategies that have typically been employed, both in matched designs. Using the example of employers, first, a researcher can send all treatment combinations to a single employer (Bertrand and Mullainathan 2004; Pager, Western, and Bonikowski 2009). Second, a researcher can randomly assign the first treatment to the employer (i.e. the cluster) and then send both the treatment and control of the second treatment to each employer (Pager 2003; Correll et al. 2007; Uggen et al. 2014). This choice often hinges on the possibility of being discovered or caught doing the audit, as sending many applications that are too similar on all other characteristics (a necessity to isolate the treatment) could arouse suspicion. But as we demonstrate in Tables 3 and 5, this choice also has efficiency implications that hinge on the expected degree of concordance, as those two treatments at the between- and within-cluster levels exhibit opposite efficiency and the interaction of the two treatments must also be considered. Thus, this decision is consequential, directly affecting a researcher's ability to detect a statistically significant effect for each treatment.

We want to emphasize a third strategy that would minimize the chance of being discovered while conducting an audit: utilizing an unmatched design when it is expected to be the more efficient design. In this approach, no single employer would be confronted with two applications that look so similar as to raise suspicion. There are certainly scenarios where suspicion could still be aroused, for example, if a researcher does not realize that two

establishments share an owner and manager. But this would occur in the matched case as well, and would likely be even more detectable as there would be multiple applications at both sites. An empirical example of this strategy is in Rivera and Tilcsik's (2016) audit of law firms, where they sent a single application to reduce the risk of discovery. Notably, however, this was likely the less efficient strategy, as the modal response was overwhelmingly for no applicant to be invited to an interview (which would likely have also been the case had a matched pair been sent). Thus, any gains in efficiency may have been more than offset by the greater likelihood of detection.

To this point, we have only described the case of two treatment/control levels. Whenever researchers want to send multiple treatments to a single experimental unit, the odds of detection typically increase. One strategy in such cases involves sending subsets of the various possible treatments to a single employer (Wright, et al. 2013; Gaddis and Ghoshal 2015b). An unmatched design, however, is even less likely to be discovered in the field. But what of efficiency? Our results above concerning the .5 concordance threshold also apply in the case of more than two treatment/control categories, whose corresponding matched statistical test is known as Cochran's *Q* (Cochran 1950). The formula we derived (Vuolo et al. 2016) from Donner and Li (1990), shows that the same size requirements for the unmatched Pearson's chi-squared test are related to a matched test via a weight that measures intraclass correlation, and is applicable regardless of the number of treatments. Thus, even if one expands the table to $2 \times m$, where m is the number of treatment categories, the preference between matched or unmatched in terms of efficiency still depends on whether the concordance is above or below .5. An unmatched design may not only be more efficient, but it may also reduce the chance of being detected while conducting an audit.

Finally, we emphasize the importance of quality randomization, as departures from randomization are even more problematic in an unmatched design. Why is this the case? Experiments are typically considered the gold standard of research for making causal claims. The randomization process renders the influence of other covariates ignorable (Quillian 2006; Pager 2007). Proper randomization in experiments, however, is demanding, even though the advantages of the method are predicated on it (Berk 2005). If the randomization process is compromised or incomplete, the result would be correlation between the treatment and observed or unobserved covariates, which would limit or altogether prevent the researcher from making the causal inferences that motivated the study. This is not the case when randomization is done well (except by random chance). While certainly not preferable to a well-conducted experiment, in the case of the matched design with incomplete randomization, researchers can fall back on fixed effects to estimate the causal influence of the treatment via the comparison of the outcomes between the two treatments at a single experimental unit (Winship and Morgan 1999; Halaby 2004). With bad randomization in an unmatched design, however, there are no post-hoc remedies to prevent the influence of covariates on the treatment effect, except classic non-causal covariate adjustment. Regardless, we emphasize the need for quality randomization. In a perfect case, sampled units should be pulled completely randomly from the population, and then randomly assigned a treatment category. For live testers, rotating the treatment among the testers is of the upmost importance so that treatments are not confounded with tester effects. Further, all treatments must be simultaneously conducted throughout the experiment, as seasonality (e.g., Schwartz and Skolnick 1962) or an exogenous shock such as a recession (e.g., Vuolo, Uggen, and Lageson, forthcoming) could alter the outcomes. Such a shock could be correlated with a given treatment, if that treatment was more likely to be assigned at a certain point in the data

collection. In the end, quality randomization should always be a priority, which would make this substantive consideration unnecessary.

4. Conclusions and Recommendations

Although matched designs are often touted for their efficiency over unmatched designs, we demonstrated that for a dichotomous outcome, this conclusion is not always justified (see Hedberg and Ayers 2015 for an argument concerning continuous outcomes in a paired *t*-test). Rather, the degree of concordance dictates whether the matched or unmatched design is more efficient. In a situation where concordance is above .5 in the population, the experimental unit itself is exerting an effect because the majority of employers gave the same response for each test, regardless of treatment or control condition. In this case, a matched design is preferable in terms of efficiency, based on the "more important" treatment that was varied within an experimental unit and its interaction with any randomly assigned cluster-level treatment. When concordance is below .5 in the population, the experimental unit is exerting a smaller effect because employers view the two applicants differently. In this case, an unmatched design is more efficient. We caution, however, that there are cases in which the unmatched design is more efficient, even when the concordance is above .5. And of course, substantive considerations are of utmost importance in creating the research design, as discussed above.

We conclude with recommendations for researchers to parse out this difficult a priori decision in the real world, building and expanding upon those in our past work (Vuolo et al. 2016). Most importantly, researchers should complete an anticipated version of Table 2 at the design stage so that they can calculate the appropriate sample size and make an informed decision between a matched and unmatched design. We recommend making several versions of

this table in order to establish expected lower and upper boundaries. As in all power calculations, this information can come from two sources. First, past studies of a similar treatment can be used. Our second and preferred recommendation is to also conduct a small pilot. We note that all sample size calculations are based only on the *proportions* in each of the four cells (and the resultant marginals in the unmatched case). Even a small pilot can assist in filling out those proportions and providing bounds.

If the calculated sample sizes for matched and unmatched designs are close or overlap to a great degree within the bounds used, we recommend taking into account the substantive considerations discussed above. When sample size calculations are close in both matched and unmatched designs, the matched design may be preferable for maintaining the possibility of estimating fixed effects (if randomization is compromised) or if twice as many experimental units for the unmatched design may exceed available elements of the population. On the other hand, if one is testing many treatment levels, the unmatched design may be preferable in the interest of not being discovered conducting the audit. Resource constraints may also make the more efficient design preferable, regardless of how close the calculation is.

The matched audit design has become very popular, seemingly becoming the norm due to a perceived efficiency gain and historical momentum. Unmatched audit designs are less common, but are beginning to appear, as shown in Table 1. In most published studies thus far, the outcome has relatively high concordance, in part because the most common response among employers (the most commonly used experimental unit) is not calling either applicant back. Thus, from an efficiency perspective, researchers will likely continue to prefer matched designs. As audits of other types of experimental units become increasingly common, however, the degree of expected concordance is likely to vary substantially (see, e.g. Lauster and

Easterbrook's (2011) audit of landlords and Wright et al.'s (2015) audit of prospective church members), with the efficiency implications we demonstrate through our hypothetical example. As audits expend considerable resources and yield important causal inferences, our results here show that the resultant decision – to match or not to match – should be a central consideration in the design of social experiments.

Acknowledgments

The illustrative empirical data used in this article come from a study conducted in partnership with the Council on Crime and Justice and supported by the JEHT Foundation and the National Institute of Justice [grant number 2007-IJ-CX-0042]. We are grateful to Laura DeMarco for the example of a landlord audit and Rob Stewart for the example of college admissions, with each coming from their respective dissertations. The R functions referenced herein, including instructions for use, are publicly available on the first author's website.

References

Berk RA (2005) Randomized experiments as the bronze standard. J Exp Criminol 1:417-433.

Bertrand M, Mullainathan S (2004) Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. Am Econ Rev 94:991-1013.

Chen D, Peace KE (2011) Clinical trial data analysis using R. CRC Press, Boca Raton, FL.

Cochran WG (1950) The comparison of percentages in matched samples. Biometrika 37:256-266.

Correll SJ, Benard S, Paik I (2007) Getting a job: is there a motherhood penalty? Am J Sociol 112:1297-1339.

Dalgaard P (2008) Introductory statistics with R, 2nd edn. Springer, New York.

Decker SH, Ortiz N, Spohn C et al (2015) Criminal stigma, race, and ethnicity: the consequences of imprisonment for employment. J Crim Just 43:108-121.

Desmond M (2016) Evicted: poverty and profit in the American city. Crown, New York.

Donner A, Li KYR (1990) The relationship between chi-square statistics from matched and unmatched analyses. J Clin Epidemiol 43:827–831.

Evans DN, Porter JR (2015) Criminal history and landlord rental decisions: a New York quasi-experiment study. J Exp Criminol 11:21-42

Gaddis SM (2015) Discrimination in the credential society: an audit study of race and college selectivity in the labor market. Soc Forces 93:1451-1479.

Gaddis SM, Ghoshal R (2015a) Arab American housing discrimination, ethnic competition, and the contact hypothesis. Ann Am Acad Polit SS 660:282-299.

Gaddis SM, Ghoshal R (2015b) Finding a roommate on craigslist: racial discrimination and residential segregation. Social Science Research Network (SSRN), Working Paper Series.

Halaby CN (2004) Panel models in sociological research: theory into practice. Annual Rev Sociol 30:507-544.

Hedberg EC, Ayers S (2015) The power of a paired t-test with a covariate." Soc Sci Res 50:277-291.

Hipes C, Lucas J, Phelan JC et al (2016) The stigma of mental illness in the labor market. Soc Sci Res 56:16-25.

Hogan B, Berry B (2011) Racial and ethnic biases in rental housing: an audit study of online apartment listings. City Community 10:351-372.

Kramer MS (1991) Clinical biostatistics: an overview. In: Troidl H, Spitzer WO, Mulder DS et al (eds) Principles and practice of research: strategies for surgical investigators, 2nd edn. Springer, New York, p 126-143.

Kugelmass H (2016) 'Sorry, I'm not accepting new patients': an audit study of access to mental health care. J Health Soc Behav 57:168-183.

Lauster N, Easterbrook A (2011) No room for new families? A field experiment measuring rental discrimination against same-sex couples and single parents. Soc Probl 58:389-409.

McNemar Q (1947) Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12:153-157.

Michel E (2016) Discrimination against queer women in the U.S. workforce: a resume audit study. Socius 2:1-13.

Pager D (2003) The Mark of a criminal record. Am J Sociol 108:937-975.

Pager D (2007) The use of field experiments for studies of employment discrimination: contributions, critiques, and directions for the future. Ann Am Acad Polit SS 609:104-133.

Pager D, Western B, Bonikowski B (2009) Discrimination in a low-wage labor market a field experiment. Am Sociol Rev 74:777-799.

Pedulla DS (2016) Penalized or protected? Gender and the consequences of nonstandard and mismatched employment histories. Am Sociol Rev 81:262-289.

Quillian L (2006) New approaches to understanding racial prejudice and discrimination. Annual Rev Sociol 32:299-328.

R Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Available via http://www.R-project.org.

Rivera LA, Tilcsik A (2016) Class advantage, commitment penalty: the gendered effect of social class signals in an elite labor market. Am Sociol Rev 81:1097-1131.

Rosner B (2011) Fundamentals of biostatistics, 7th edn. Brooks/Cole, Boston.

Schwartz RD, Skolnick JH (1962) Two studies of legal stigma. Soc Problems 10:133-142.

Shih WJ, Aisner J (2016) Statistical design and analysis of clinical trials: principles and methods. CRC Press, New York.

Stewart R, Uggen C (2016) A modified experimental audit of criminal records and college admissions. Paper presented at the American Society of Criminology Meetings, New Orleans, 17 November 2016.

Tilcsik A (2011) Pride and prejudice: employment discrimination against openly gay men in the United States." Am J Sociol 117:586-626.

Uggen C, Vuolo M, Lageson S et al (2014) The edge of stigma: an experimental audit of the effects of low-level criminal records on employment. Criminology 52:627-654.

Vuolo M, Uggen C, Lageson S (2016) Statistical power in experimental audit studies: cautions and calculations for matched tests with nominal outcomes. Soc Method Res 45:260-303.

Vuolo M, Uggen C, Lageson S (Forthcoming) Race, recession, and social closure in the low wage labor market: experimental and observational evidence. Res Sociol Work.

Wallace M, Wright BRE, Hyde A (2014) Religious affiliation and hiring discrimination in the American south: a field experiment. Soc Currents 1:189-207.

Widner D, Chicoine S (2011) "It's all in the name: employment discrimination against Arab Americans." Sociol Forum 26:806-823.

Winship C, Morgan SL (1999) The estimation of causal effects from observational data. Annual Rev Sociol 659-706.

Wright BRE, Wallace M, Bailey J et al (2013) Religious affiliation and hiring discrimination in New England: a field experiment. Res Soc Stratif Mobil 34:111-126.

Wright BRE, Wallace M, Wisnesky AS et al (2015) Religion, race, and discrimination: a field experiment of how American churches welcome newcomers. J Sci Stud Relig 54:185-204.

Table 1: Examples of Audits in Sociology and Criminal Justice by Matching Design

Author (Year)	Journal	Treatment(s)	Unit	Matched?
Vuolo, Uggen, & Lageson (forthcoming)	*Res Sociol Work*	Race	Employer	No
Mishel (2016)	*Socius*	Sexual identity	Employer	Yes
Rivera & Tilcsik (2016)	*Am Sociol Rev*	Class	Employer	No
Pedulla (2016)	*Am Sociol Rev*	Employment history, gender	Employer	Yes
Hipes, et al. (2016)	*Soc Sci Res*	Mental illness	Employer	No
Kugelmass (2016)	*J Health Soc Behav*	Race, class, gender	Psychotherapists	Yes
Gaddis (2015)	*Soc Forces*	Race, academics	Employer	Yes
Gaddis & Ghoshal (2015a)	*Ann Am Acad Polit SS*	Race/ethnicity	Roommates	Yes
Wright, et al. (2015)	*J Sci Stud Relig*	Race/ethnicity	Churches	No
Decker, et al. (2015)	*J Crim Just*	Race/ethnicity, prison	Employer	Yes
Evans & Porter (2015)	*J Exp Crim*	Criminal record, gender	Landlords	Yes
Uggen, et al. (2014)	*Criminology*	Misdemeanor arrest, race	Employer	Yes
Wallace, Wright, & Hyde (2014)	*Soc Currents*	Religion	Employer	Yes
Wright, et al. (2013)	*Res Soc Stratif Mobil*	Religion	Employer	Yes
Widner & Chicoine (2011)	*Sociol Forum*	Arab ethnicity	Employer	Yes
Lauster & Easterbrook (2011)	*Soc Problems*	Sexual orientation, parenthood	Landlords	No
Hogan & Barry (2011)	*City Community*	Race	Landlords	Yes
Tilcsik (2011)	*Am J Sociol*	Sexual orientation	Employer	Yes
Pager, Western, & Bonikowski (2009)	*Am Sociol Rev*	Felony, race/ethnicity	Employer	Yes
Correll et al. (2007)	*Am J Sociol*	Parenthood, gender	Employer	Yes
Pager (2003)	*Am J Sociol*	Felony, race	Employer	Yes

Table 2: Audit Notation

		Treatment		
		Affirmative response	Negative/No response	Total
Control	Affirmative response	n_{11} p_{11}	n_{10} p_{10}	n_{1+} p_{1+}
	Negative/No response	n_{01} p_{01}	n_{00} p_{00}	n_{0+} p_{0+}
	Total	n_{+1} p_{+1}	n_{+0} p_{+0}	n $p_{++} = 1$

Table 3: Distribution of Callbacks by Criminal Record for Each Paired Employer Audit (Uggen et al. 2014)

Total

		Misdemeanor Arrest		
		Callback	No Callback	Total
No Misdemeanor Arrest	Callback	$n_{11} = 60$ $p_{11} = .200$	$n_{10} = 39$ $p_{10} = .130$	$n_{1+} = 99$ $p_{1+} = .330$
	No Callback	$n_{01} = 27$ $p_{01} = .090$	$n_{00} = 174$ $p_{00} = .580$	$n_{0+} = 201$ $p_{0+} = .670$
	Total	$n_{+1} = 87$ $p_{+1} = .290$	$n_{+0} = 213$ $p_{+0} = .710$	$n = 300$ $p_{++} = 1$

African American

		Misdemeanor Arrest		
		Callback	No Callback	Total
No Misdemeanor Arrest	Callback	$n_{11} = 24$ $p_{11} = .157$	$n_{10} = 18$ $p_{10} = .118$	$n_{1+} = 42$ $p_{1+} = .275$
	No Callback	$n_{01} = 12$ $p_{01} = .078$	$n_{00} = 99$ $p_{00} = .647$	$n_{0+} = 111$ $p_{0+} = .725$
	Total	$n_{+1} = 36$ $p_{+1} = .235$	$n_{+0} = 117$ $p_{+0} = .765$	$n = 153$ $p_{++} = 1$

White

		Misdemeanor Arrest		
		Callback	No Callback	Total
No Misdemeanor Arrest	Callback	$n_{11} = 36$ $p_{11} = .245$	$n_{10} = 21$ $p_{10} = .143$	$n_{1+} = 57$ $p_{1+} = .388$
	No Callback	$n_{01} = 15$ $p_{01} = .102$	$n_{00} = 75$ $p_{00} = .510$	$n_{0+} = 90$ $p_{0+} = .612$
	Total	$n_{+1} = 51$ $p_{+1} = .347$	$n_{+0} = 96$ $p_{+0} = .653$	$n = 147$ $p_{++} = 1$

Table 4: Logistic model simulations for case of high concordance

	Unmatched	Matched cluster correction		Matched hierarchical model	
	Simulated	Actual	Simulated	Actual	Simulated
Model 1					
Record	-0.186	-0.187	-0.187	-0.364	-0.368
	(0.215)	(0.127)	**(0.126)**	(0.249)	(0.254)
Model 2					
Record	-0.190	-0.190	-0.190	-0.364	-0.370
	(0.219)	(0.129)	**(0.129)**	(0.249)	(0.254)
White	0.532***	0.530*	0.536*	1.018*	1.028*
	(0.130)	(0.218)	(0.220)	(0.422)	(0.429)
Model 3					
Record	-0.210	-0.207	-0.209	-0.390	-0.398
	(0.326)	(0.189)	**(0.192)**	(0.363)	(0.371)
White	0.519*	0.515*	0.520*	0.995*	1.013*
	(0.252)	(0.248)	(0.250)	(0.483)	(0.494)
Record * White	0.033	0.031	0.033	0.048	0.051
	(0.442)	(0.258)	**(0.262)**	(0.495)	(0.510)

*$p < .05$, **$p < .01$, ***$p < .001$

Note: We drew 100,000 simulations of a dataset with 300 observations. These data are drawn from a dataset of 600, represented by the "Actual" model. Lowest standard error for a coefficient is bolded.

Table 5: Hypothetical Distribution with Lower Concordance (.25) of Callbacks by Mention of Family for Each Paired Audit of Landlords

Total

		Family mentioned		
		Callback	No Callback	Total
No family mentioned	Callback	$n_{11} = 19$ $p_{11} = .063$	$n_{10} = 135$ $p_{10} = .450$	$n_{1+} = 154$ $p_{1+} = .513$
	No Callback	$n_{01} = 90$ $p_{01} = .300$	$n_{00} = 56$ $p_{00} = .187$	$n_{0+} = 146$ $p_{0+} = .487$
	Total	$n_{+1} = 109$ $p_{+1} = .363$	$n_{+0} = 191$ $p_{+0} = .637$	$n = 300$ $p_{++} = 1$

African American

		Family mentioned		
		Callback	No Callback	Total
No family mentioned	Callback	$n_{11} = 5$ $p_{11} = .033$	$n_{10} = 62$ $p_{10} = .413$	$n_{1+} = 67$ $p_{1+} = .447$
	No Callback	$n_{01} = 43$ $p_{01} = .287$	$n_{00} = 40$ $p_{00} = .267$	$n_{0+} = 83$ $p_{0+} = .553$
	Total	$n_{+1} = 48$ $p_{+1} = .320$	$n_{+0} = 102$ $p_{+0} = .680$	$n = 150$ $p_{++} = 1$

White

		Family mentioned		
		Callback	No Callback	Total
No family mentioned	Callback	$n_{11} = 14$ $p_{11} = .093$	$n_{10} = 73$ $p_{10} = .487$	$n_{1+} = 87$ $p_{1+} = .580$
	No Callback	$n_{01} = 47$ $p_{01} = .313$	$n_{00} = 16$ $p_{00} = .107$	$n_{0+} = 63$ $p_{0+} = .420$
	Total	$n_{+1} = 61$ $p_{+1} = .407$	$n_{+0} = 89$ $p_{+0} = .593$	$n = 150$ $p_{++} = 1$

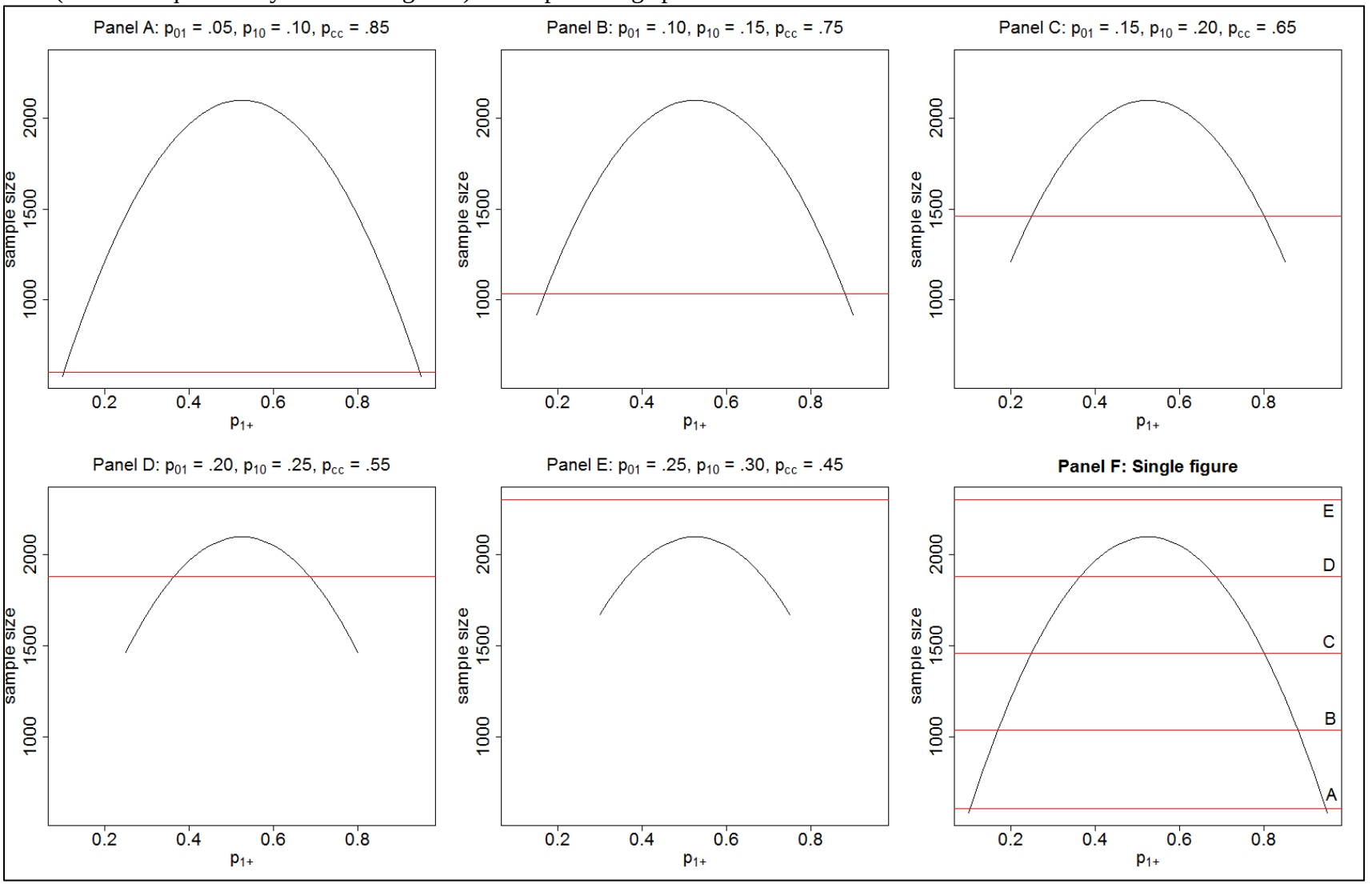
Table 6: Logistic model simulations for case of low concordance

	Unmatched	Matched cluster correction		Matched hierarchical model	
	Simulated	Actual	Simulated	Actual	Simulated
Model 1					
Family	-0.617***	-0.614**	-0.616**	-0.614***	-0.616**
	(0.114)	(0.205)	(0.205)	(0.167)	(0.204)
Model 2					
Family	-0.627***	-0.622**	-0.626**	-0.622**	-0.624**
	(0.116)	(0.208)	(0.208)	(0.168)	(0.208)
White	0.463*	0.459***	0.462***	0.459**	0.462***
	(0.210)	(0.115)	**(0.116)**	(0.168)	(0.116)
Model 3					
Record	-0.545***	-0.540	-0.544	-0.540*	-0.546
	(0.176)	(0.292)	(0.294)	(0.240)	(0.294)
White	-0.541*	0.537*	0.539*	0.537*	0.539*
	(0.235)	(0.234)	(0.235)	(0.233)	(0.235)
Family * White	-0.150	-0.161	-0.161	-0.161	-0.159
	(0.232)	(0.416)	(0.419)	(0.336)	(0.418)

*$p < .05$, **$p < .01$, ***$p < .001$

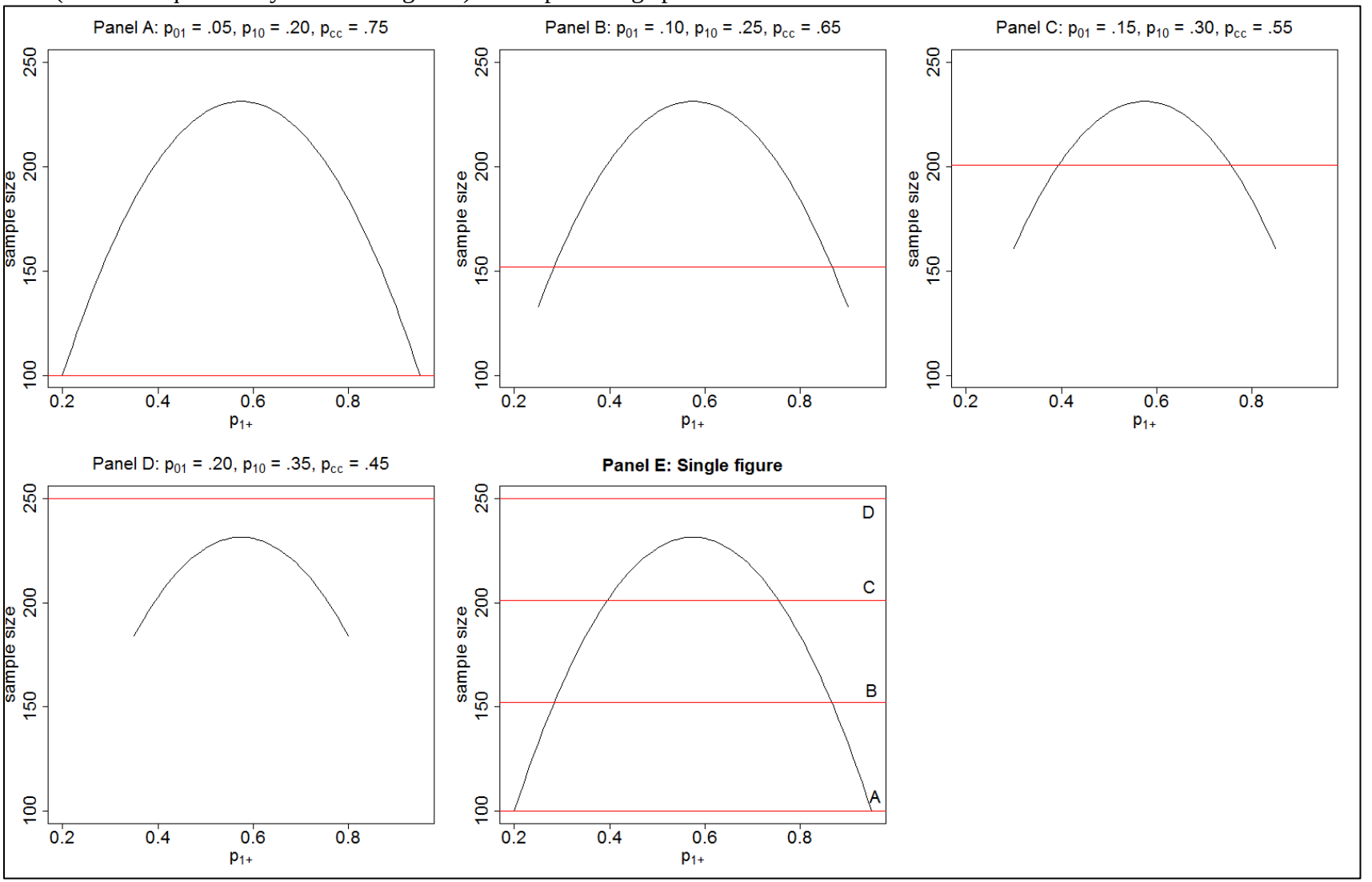
Note: We drew 100,000 simulations of a dataset with 300 observations. These data are drawn from a dataset of 600, represented by the "Actual" model. Lowest standard error for a coefficient is bolded.

Figure 1: Sample size requirements for unmatched test of proportions vs. matched McNemar's Test for a difference in the concordant cells (and thus equivalently for the marginals) of .05 percentage points



Note: All calculations are for power of .9 and $\alpha = .05$. Horizontal lines represent sample size requirement for matched design. The curve represents the sample size requirement for the unmatched design.

Figure 2: Sample size requirements for unmatched test of proportions vs. matched McNemar's Test for a difference in the concordant cells (and thus equivalently for the marginals) of .15 percentage points



Note: All calculations are for power of .9 and $\alpha = .05$. Horizontal lines represent sample size requirement for matched design. The curve represents the sample size requirement for the unmatched design.

Figure 3: Simulated Sampling Distributions for Table 4, Model 2

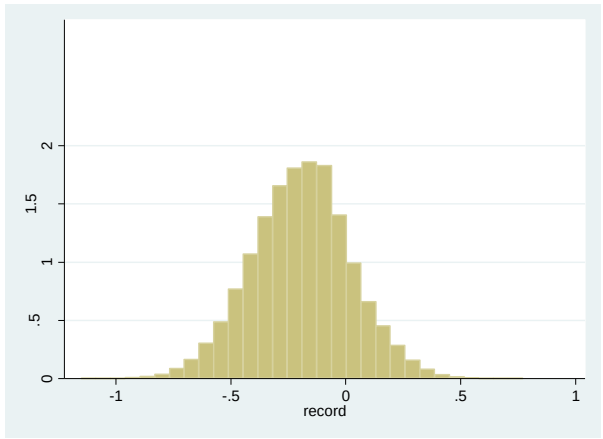
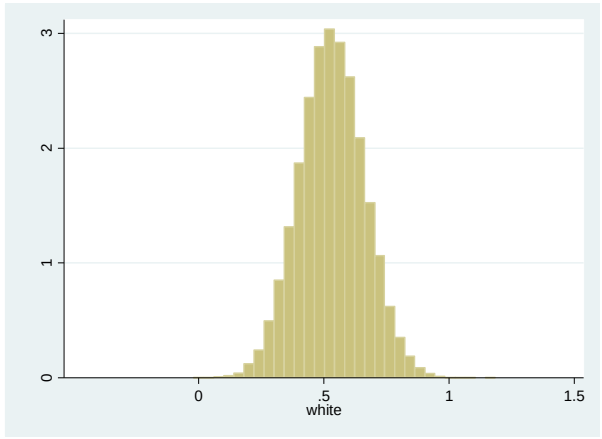
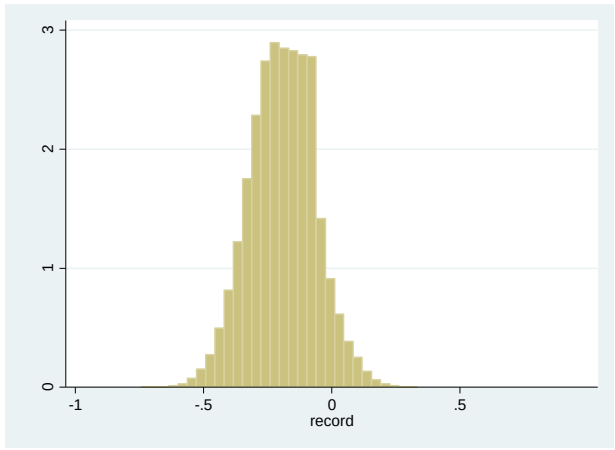
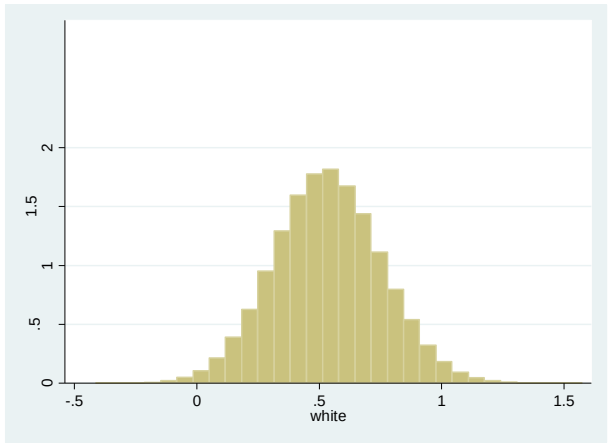
	Record	White
Unmatched		
Matched Cluster corrected		

Figure 4: Simulated Sampling Distributions for Table 4, Model 3

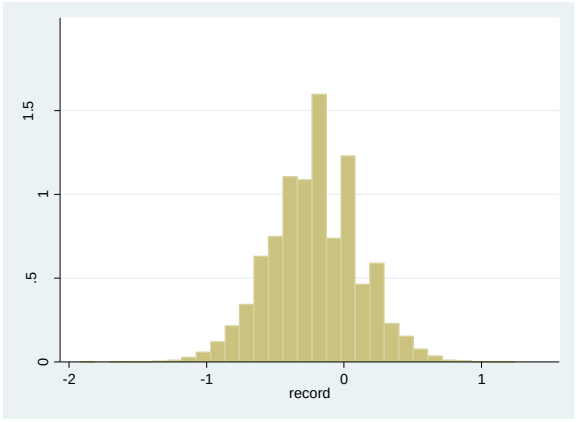
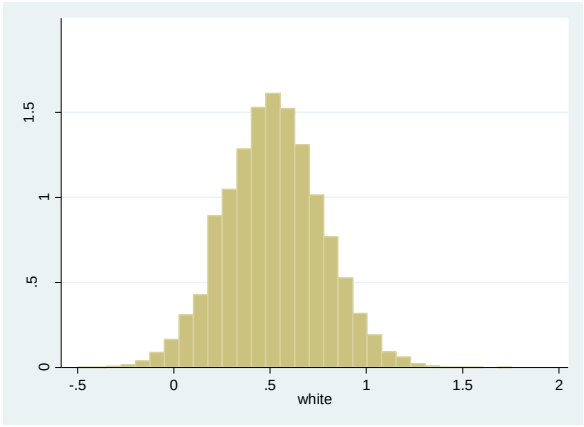
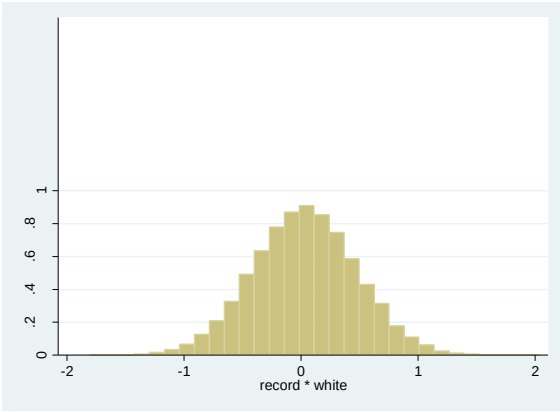
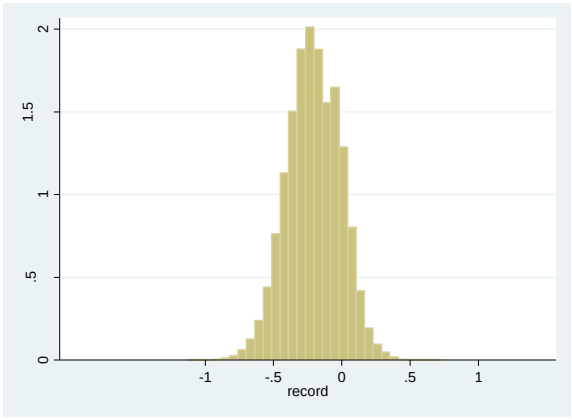
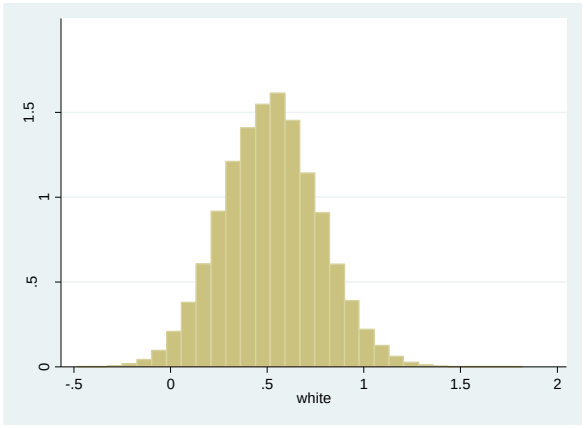
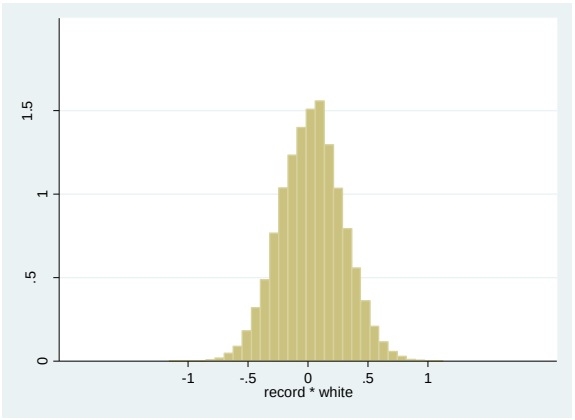
	Record	White	Interaction
Unmatched			
Matched Cluster corrected			

Figure 5: Simulated Sampling Distributions for Table 6, Model 2

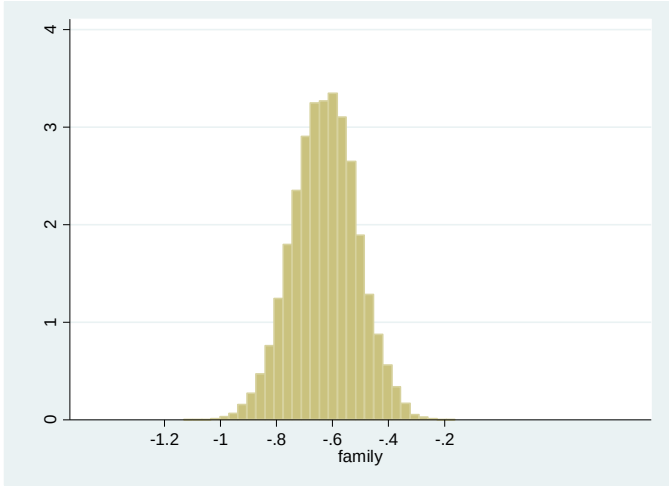
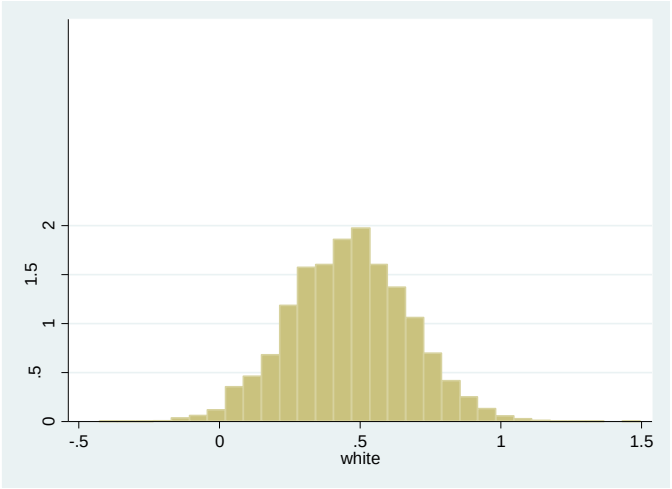
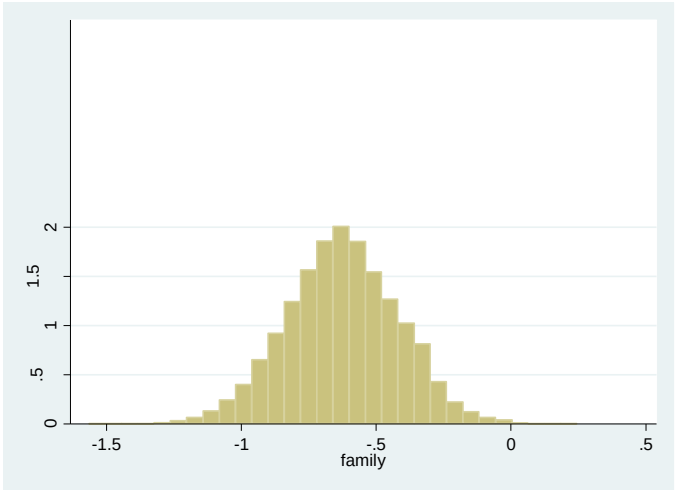
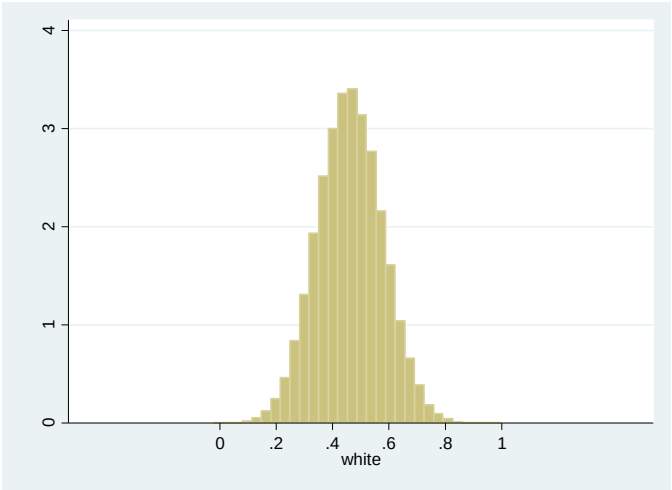
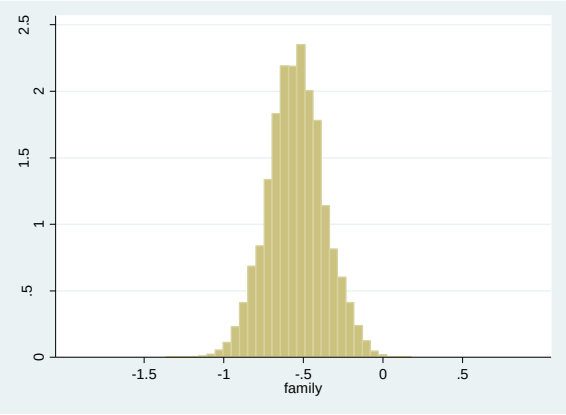
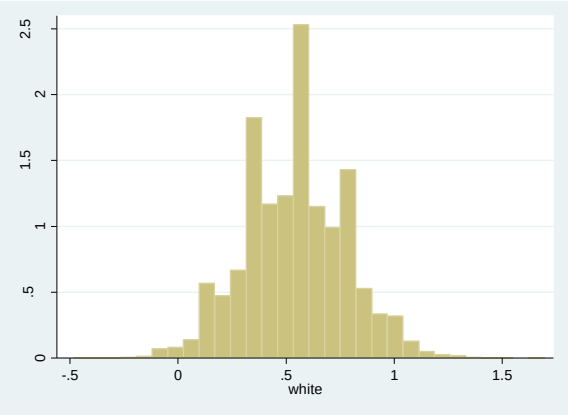
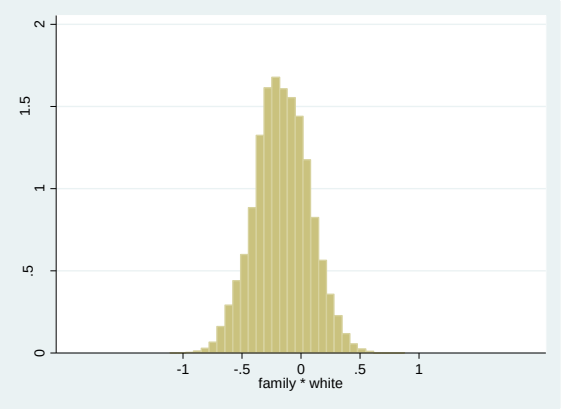
	Family	White
Unmatched		
Matched Cluster corrected		

Figure 6: Simulated Sampling Distributions for Table 6, Model 3

	Family	White	Interaction
Unmatched			
Matched Cluster corrected