Sociological Methods & Research
1-44

DOI: 10.1177/0049124115570066
smr.sagepub.com

Statistical Power in Experimental Audit Studies: Cautions and Calculations for Matched Tests With Nominal Outcomes

Mike Vuolo[1], Christopher Uggen[2], and Sarah Lageson[2]

Abstract

Given their capacity to identify causal relationships, experimental audit studies have grown increasingly popular in the social sciences. Typically, investigators send fictitious auditors who differ by a key factor (e.g., race) to particular experimental units (e.g., employers) and then compare treatment and control groups on a dichotomous outcome (e.g., hiring). In such scenarios, an important design consideration is the power to detect a certain magnitude difference between the groups. But power calculations are not straightforward in standard matched tests for dichotomous outcomes. Given the paired nature of the data, the number of pairs in the concordant cells (when neither or both auditor receives a positive response) contributes to the power, which is lower as the sum of the discordant proportions approaches one. Because these quantities are difficult to determine a priori, researchers must exercise particular care in experimental design. We here

[1] Purdue University, West Lafayette, IN, USA
[2] University of Minnesota, Minneapolis, MN, USA

Corresponding Author:
Mike Vuolo, Department of Sociology, Purdue University, 700 W State St., West Lafayette, IN 47907, USA.
Email: mvuolo@purdue.edu

present sample size and power calculations for McNemar's test using empirical data from an audit study on misdemeanor arrest records and employability. We then provide formulas and examples for cases involving more than two treatments (Cochran's *Q* test) and nominal outcomes (Stuart–Maxwell test). We conclude with concrete recommendations concerning power and sample size for researchers designing and presenting matched audit studies.

Keywords

sample size, power, experiments, audit studies, McNemar's test, Cochran's *Q* test, Stuart–Maxwell test

Introduction

Issues of statistical power are central in determining the proper sample size to detect effects of a given magnitude. In experimental audit studies, however, this important design decision is rarely discussed, in part because power calculations are not straightforward. In light of the recent surge in audit and correspondence studies since Pager (2003), it is an opportune moment to address this question. Using matching and random selection, audit studies are an experimental method using real-world contexts to uncover causal mechanisms behind social phenomena, often discrimination (Pager 2007:109). Although there are excellent guides on how to design and implement a matched audit study (e.g., Pager's 2007 appendix), questions of statistical power and sample size have yet to be addressed in the social science literature.

This article details the unique challenges of calculating power for audit studies and correspondence tests, which share characteristics such as repeated observations and nominal outcomes. 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 percent) that a statistically significant effect ($p < .05$) will be detected in a given sample? Unfortunately, for even the simplest paired design with a dichotomous outcome, the calculation of power depends on more than the simple magnitude difference. First, given the paired nature of the data, the number of total pairs in both concordant cells (i.e., both testers receive the same outcome) compared to the number of total pairs in both discordant cells (i.e., the testers each receive a different outcome) contributes

to the power. Second, as a nominal outcome, power is lower as the sum of the discordant proportions approaches 1, even for the same magnitude difference. Since these numbers are difficult to determine a priori, power calculations are challenging in the design phases of such studies. Moreover, these calculations become even more challenging as the number of either treatments or outcomes is increased.

The time has come to squarely address the challenge of power calculations for such field experiments. Although medical researchers have to some extent explicated such challenges in clinical trials (e.g., Lachenbruch 1992; Royston 1993), the distinctive audit studies mounted in the social sciences warrant a specific examination of power and sample size in field experiments. First, the exchange in the medical literature has only addressed the case of sample size challenges for a 2×2 table, which we expand to a general table. Second, the medical field approaches matched designs from the case-control perspective. Thus, when moving beyond 2×2, this typically means expanding the *response* to three categories rather than including an additional *treatment*. Here, given the typical social science audit setup, we approach the question from both perspectives. Further, much of the sample size literature for clustered tables involves randomization of the treatment *to clusters* (see, e.g., Donner 1992); that is, the entire experimental unit (e.g., siblings within a family) is assigned either the treatment or the control. Audit studies, on the other hand, test both the control and the treatment(s) at each randomly selected experimental unit. Third, field experiments pose different challenges than clinical experiments, resulting in different and distinct recommendations, which follow in our discussion section. Fourth, an empirical exercise for calculating sample size and power for an audit study has yet to be conducted, allowing social scientists to draw more relevant parallels than the example of a clinical trial. Most importantly, no article has yet to collect the statistical tests and their sample size calculations for such designs, which are each a case of the Generalized Cochran–Mantel–Haenszel (CMH) test, or to provide formulas and functions for the increasingly popular case of more than two repeated measures (Cochran's Q test).

This article thus proceeds as follows. We begin by reviewing audit studies in the social sciences, concentrating on outcomes, treatments, and sample sizes. Second, we describe the appropriate statistical tests for the paired audit design, McNemar's test, and the formulas for sample size and power. We then show calculations for sample size, demonstrating the inherent difficulties in determining sample sizes for this particular design and apply these calculations for McNemar's test to an empirical example from a 2×2 audit study on arrest records and employability, further emphasizing the

challenges of determining sample size and comparing our a priori expectations with the actual results. Third, we introduce tests and sample size calculations for situations in which the number of outcomes or treatment categories is greater than two and describe how this compounds the issue of sample size selection, including a hypothetical example to illustrate calculations for each case. Fourth, we briefly discuss extensions of the bivariate case, such as linear modeling, covariates, and additional stratification variables. Finally, in light of these calculations and cautions, we conclude with concrete recommendations for social scientists planning audit studies.

The Current Landscape of Audit Studies in the Social Sciences

Experimental audit studies have become an increasingly important methodological approach in disciplines such as sociology, economics, and criminology. We consider both audit tests and correspondence tests together. The most important distinction between these is that the former sends live testers to conduct audits, while the latter sends entirely fictional applications without a live tester. For simplicity, we use the word "audits" throughout this article, as the statistical methodology and implications for sample size are identical, though we caution that important differences exist (Pager 2007). We summarize some of the recent audit and correspondence studies conducted in the social sciences in Table 1.[1]

Three points are immediately apparent from the table. First, sample size varies greatly, as indicated by the number of sites. Some studies are based on fewer than 100 experimental units (Neumark, Bank, and Van Nort 1996), while others exceed 3,000 (Yinger 1991). Second, the treatments now being tested extend far beyond racial discrimination to include factors such as religious affiliation (Wright et al. 2013), sexual orientation (Tilcsik 2011), and parenthood status (Correll, Benard, and Paik 2007). Finally, without regard to discipline, almost all of these studies are extremely well cited. Based on Google Scholar citation data for August 2014, Bertrand and Mullainathan (2004) have been referenced over 1,600 times, while Pager (2003) has been cited over 1,000 times. Moreover, a preponderance of these studies are published in flagship journals, with some of them ranking among the most cited social science articles within the period since their publication.

Although the importance of audit studies is clearly reflected in their visibility and influence, the question of appropriate sample size has yet to be explored. There are at least four good reasons to do so. First, experimental audits have far greater capacity to reveal causal mechanisms than do

Table 1. Recent Matched Audit Studies in the Social Sciences.

Citation	Unit	Treatment	Sites	Testers	Location	Outcome	Type	Cites
1. Wright et al. (2013)	Employers	Religious affiliation	1,600	4	New England	Callback	Correspondence	1
2. Tilcsik (2011)	Employers	Sexual orientation	1,796	2	Seven states	Invitation to interview	Correspondence	49
3. Pager, Western, and Bonikowski (2009)	Employers	Felony, race	171/169	3	New York	Callback/offer	Audit	237
4. Drydakis (2009)	Employers	Sexual orientation	1,714	2	Athens, Greece	Callback, wage	Correspondence	64
5. Lahey (2008)	Employers	Age	3,996	2	Boston, St. Petersburg	Positive response, interview	Correspondence	101
6. Correll et al. (2007)	Employers	Parenthood	638	2	Northeast City	Callback	Correspondence	584
7. Bertrand and Mullainathan (2004)	Employers	Race, skill	1323	4	Chicago, Boston	Callback	Correspondence	1,664
8. Pager (2003)	Employers	Felony	350	2	Milwaukee	Callback	Audit	1,069
9. Weichselbaumer (2003)	Employers	Sexual orientation	613	2	Austria	Invitation	Correspondence	143
10. Bendick, Brown, and Wall (1999)	Employers	Age	102	2	Washington, DC	Favorable response	Audit	92
11. Bendick, Jackson, and Romero (1997)	Employers	Age	775	2	United States	Positive response	Correspondence	57
12. Esmail and Everington (1997)	Medical schools	Ethnicity	50	2	Great Britain	Callback, shortlist	Correspondence	47
13. Neumark, Bank, and Van Nort (1996)	Employer	Gender	65	2	Philadelphia	Callback	Audit	355

(continued)

Table 1. (continued)

Citation	Unit	Treatment	Sites	Testers	Location	Outcome	Type	Cites
14. Ayres and Siegelman (1995)	Car dealers	Race, gender	153	2	Chicago	Offer price; accept counter	Audit	452
15. Bendick, Jackson, and Reinoso (1994)	Employers	Race	149	2	Washington, DC	Job offer	Audit	154
16. Kenney and Wissoker (1994)	Employer	Ethnicity	302	2	Chicago, San Diego	Callback, 3 levels	Audit	121
17. Esmail and Everington (1993)	Medical schools	Ethnicity	23	2	Great Britain	Callback, shortlisted	Correspondence	153
18. Turner and Mikelsons (1992)	Housing	Race	1,081/1,076/ 801/787	2	25 U.S. Metro Areas	3 Stages	Audit	27
19. Turner, Fix, and Struyk (1991)	Employers	Race	418	2	Chicago; Washington, DC	Job offer	Audit	218
20. Ayres (1991)	Car dealers	Race, gender	90	2	Chicago	Offer price; accept counter	Audit	645
21. Bendick et al. (1991)	Employers	Race	468	2	Washington, DC	Positive response	Correspondence	47
22. Riach and Rich (1991)	Employer	Race, ethnicity	519/462	2	Victoria, Australia	Interview offer	Correspondence	82
23. Yinger (1991)	Housing	Race	1,081/1,076/ 801/787	2	25 US Metros	3 stages	Audit	26
24. Riach and Rich (1987)	Employer	Gender	991	2	Victoria, Australia	Interview offer	Correspondence	50
25. Yinger (1986)	Housing	Race	156	2	Boston	Units offered	Audit	369
26. Feins and Bratt (1983)	Housing	Race	274	2	Boston	Available units	Audit	34

alternatives such as covariate adjustment analysis of survey data through the process of randomization, which for audits is accomplished via random selection and matching (Pager 2007; Quillian 2006). These research endeavors are thus of great scientific and policy importance. In fact, randomized experiments are often considered the "gold standard" of research (e.g., Sherman et al. 1998). Further, audits are capable of uncovering dimensions of social phenomena, such as discrimination, that are difficult to study otherwise, as expressed behavior in surveys and interviews is demonstrably different from that of actual behavior in audits (Lageson, Vuolo, and Uggen 2014; Pager and Quillian 2005). Thus, maximizing the chance of detecting a significant effect in audit studies should take a central role in the design phase.

The remaining three reasons are closely related to cost, funding, and feasibility. Following second then, audit studies are both time and resource intensive (see Appendix in Pager 2007), particularly those utilizing in-person data collection rather than correspondence-based tests. Although the latter cost substantially less because no live testers must be paid, research personnel are still needed in both cases to conduct a range of labor-intensive tasks (such as sampling, random assignment, quality control, and tracking responses). When researchers overestimate the number of sites required to detect a practically and statistically significant effect, valuable resources are wasted, especially for in-person experiments. Third, in the absence of a reasonable guide to power calculations, researchers will err on the side of caution, exhausting resources that could otherwise be deployed to expand the study to additional treatments or outcomes (e.g., adding a Latino group to a proposed test of discrimination against African Americans relative to Whites, adding women to a proposed study of discrimination against those with criminal records, and expanding response categories in hiring to distinguish between an interview offer and a job offer). Thus, proper sample size calculations for both additional treatments and outcomes could increase the scientific yield of studies without incurring additional costs.

Both of these lead to the final important reason for sample size calculations: from the perspective of proposal reviewers and granting agencies, there are currently few standards to evaluate the appropriateness of the proposed sample size and budget in audit studies. Particularly in these times of scarce and declining funding, there are tremendous opportunity costs for either overfunding (in terms of other studies that are not funded) or underfunding (in terms of the funded audit's capacity to advance knowledge). Demonstrating the ability for increased scientific yield and appropriately allocated resources will both improve research proposals and assist in accurately appraising them for funding. Given the costs in time, resources, and

money and the related potential to expand a study, maximizing power through careful sample size calculations is paramount.

Although power calculations are important for all of these reasons, the issues of matching and nominal outcomes complicate their application in experimental audit studies. We begin by introducing the appropriate statistical test for these designs.

The Generalized Cochran-Mantel-Haenszel Test

With dichotomous experimental outcomes, as in the social science audit studies outlined previously, the choice of an appropriate statistical test is not as straightforward as it is for a continuous experimental response variable, for which the most appropriate statistical methods are variants of matched-pairs *t*-tests. Specifically, these experiments conform to a completely randomized block design (Pinheiro and Bates 2004), but with a nominal outcome in the case of an audit study. The block makes audit and correspondence studies unique relative to other field experimental designs, where each experimental unit only receives either a treatment or a control (e.g., the well-known field experiments testing mandatory arrest in police calls for spousal abuse; Berk et al. 1992; Sherman and Berk 1984). The block then is the experimental unit that is randomly selected, with each unit having repeated measures of both the treatment(s) and the control. To broadly apply the discussion that follows, we refer to the object whose response is being tested as the *experimental unit*, the dichotomous outcome coded 1 as an *affirmative response*, *tester* as the individual observations or cases presented to the experimental unit, *treatment* as the presence of the condition being tested, and *control* as the absence of the condition. For example, in the empirical example that follows for a dichotomous outcome with a single treatment, we tested whether employers (experimental unit) called back (affirmative response) either of two job applicants (testers), when one presented an arrest record (treatment) and the other presented a clean record (control).

We next describe the larger family of statistics to which audit studies conform. The appropriate statistical tests for matched experiments with nominal outcomes fall under a larger family of statistics known as generalized Cochran-Mantel-Haenszel (CMH) tests for $O \times R \times S$ tables, where $O =$ number of outcome response categories, $R =$ number of repeated measures (i.e., treatments plus controls), and $S =$ number of strata (Birch 1965; Landis, Heyman, and Koch 1978; Mantel and Byar 1978).[2] For a matched design, $S = N$; that is, a matched design is a generalized CMH test where each

experimental unit is its own stratum (Agresti 2002:413–14, 458–59).[3] In determining sample size, it is this latter value that we seek to compute, as O and R are typically predetermined based on the research question. Table 2 summarizes the various tests in this family based on the values of O and R whose sample size selection we will discuss. As shown in the table, analyzing paired data with two repeated measures and a dichotomous outcome, the generalized CMH reduces to McNemar's (1947) test. An example of such a design is the empirical example described previously. When there are multiple treatments and a dichotomous outcome, the statistic is known as Cochran's Q (Cochran 1950). For example, a researcher could send testers of three racial categories to the same employers and measure whether each receives a callback. Finally, with two repeated measures and a nominal outcome with more than two categories, the Stuart–Maxwell test is appropriate (Stuart 1955). Although unexplored up to this point, a hypothetical example would include sending two testers of differing racial categories to the same employer and measuring three different responses, such as receiving no callback, receiving a callback for the position advertised, or receiving a callback for a lesser position (see, e.g., Pager, Western, and Bonikowski 2009). We will first address the more straightforward case of two repeated measures with a dichotomous outcome, which itself presents considerable sample size selection challenges, before addressing the case of more than two repeated measures or outcome categories.

McNemar's Test: One Treatment and a Dichotomous Outcome

Formulas and Notation for McNemar's Test

The simplest case of a 2×2 paired table conforms to McNemar's test, with the notation shown in Table 3 following Agresti (2002:410–11). If we consider the main treatment effect of interest in most of the articles outlined in Table 1 (see "Testers" column), they conform to this table.[4] In Table 3, π_{ab} denotes the population probability of outcome a for the first tester and outcome b for the other tester for the same experimental unit (i.e., different or discordant outcomes). n_{ab} represents the count of the number of *pairs* in each cell, with sample proportion equal to $p_{ab} = n_{ab}/n$. In what follows, we denote an affirmative response as 1 and a negative response as 0 in the subscripts. The first subscript represents the outcome for the control tester, while the second subscript represents the outcome for the treatment tester for the same experimental unit. McNemar's test assesses the hypothesis of marginal homogeneity or H_0: $\pi_{1+} = \pi_{+1}$. This is equivalent to testing equality between

Table 2. Types of Generalized Cochran-Mantel-Haenszel Tests.

	$O \times R$	Example	Empirical examples from Table 1
Generalized CMH test	$O \times R$		
McNemar's test	2×2	Two applicants sent to same employers, one with a misdemeanor arrest, and callback measured as outcome	2, 4-6, 8-26
Cochran's Q test	$2 \times R$	Applicants of three differing racial categories (e.g. White, Black, Latino) sent to same employers, and callback measured as outcome	1, 3, 7
Stuart–Maxwell test	$O \times 2$	Two applicants of differing racial categories sent to same employers and the measured outcome is no callback, callback but channeled downward to lower position, and callback for position to which applied	None

Note: A Generalized CMH Test is typically of the form $O \times R \times S$, where S is the number of strata. For a matched design, each stratum represents a unique experimental unit, such that $S = N$. O = number of outcome response categories; R = number of repeated measures (i.e., treatments + controls).

the cells in which testers had different outcomes (i.e., H_0: $\pi_{10} = \pi_{01}$), or that the difference between the two discordant proportions is zero in the population. We can easily prove this result by $\pi_{1+} - \pi_{+1} = (\pi_{11} + \pi_{10}) - (\pi_{11} + \pi_{01}) = \pi_{10} - \pi_{01}$. That is, the common occurrence of an affirmative response at the same experimental unit is netted out of the marginal distribution, which is what makes the matched case unique from the independent case. The ratio $z = (p_{10} - p_{01})/\hat{\sigma}_{p_{1+}-p_{+1}}$ is a Wald test statistic. Under the null hypothesis, the variance is estimated as:

$$\hat{\sigma}_{p_{1+}-p_{+1}} = \frac{p_{10} + p_{01}}{n} = \frac{n_{10} + n_{01}}{n^2}. \tag{1}$$

Thus, inserting the variance into the Wald ratio, the test statistic simplifies to the following:

$$\chi_1^2 = \frac{(n_{10} - n_{01})^2}{n_{10} + n_{01}}. \tag{2}$$

Table 3. Notation for McNemar's Test for Matched Proportions.

		Treatment		
		Affirmative Response	Negative Response	Total
Control	Affirmative response	Population proportion: π_{11} Sample proportion: p_{11} Sample cell size: n_{11}	Population proportion: π_{10} Sample proportion: p_{10} Sample cell size: n_{10}	Population proportion: π_{1+} Sample proportion: p_{1+} Sample cell size: n_{1+}
	Negative response	Population proportion: π_{01} Sample proportion: p_{01} Sample cell size: n_{01}	Population proportion: π_{00} Sample proportion: p_{00} Sample cell size: n_{00}	Population proportion: π_{0+} Sample proportion: p_{0+} Sample cell size: n_{0+}
	Total	Population proportion: π_{+1} Sample proportion: p_{+1} Sample cell size: n_{+1}	Population proportion: π_{+0} Sample proportion: p_{+0} Sample cell size: n_{+0}	Population proportion: $\pi_{++} = 1$ Sample proportion: $p_{++} = 1$ Sample cell size: n

Note that the McNemar test statistic depends only on cases classified in different categories (i.e., the discordant cells) for the two matched observations. Of course, all cases contribute to inferences about how much π_{10} and π_{01} differ because the total proportion in the two concordant cells affects the total proportion in the two discordant cells, as they must add to 1. Yet, for this experimental design and test, the breakdown between the two concordant cells is irrelevant. In other words, the results are identical regardless of the split between the cells where both testers simultaneously receive affirmative or negative responses (i.e., only $p_{11} + p_{00}$ is relevant, but not either term individually). This fact becomes important in power and sample size calculations, as two relationships determine the values: (1) the total discordant proportion ($p_{10} + p_{01}$) relative to the total concordant proportion ($p_{11} + p_{00}$) and (2) the relative proportion in the two discordant cells (p_{10} and p_{01}).

Following Rosner (2011:384–86), we provide the equations for power and sample size for McNemar's test. Let $p_{c/DD} = \left(\frac{n_{10}}{n_{10}+n_{01}}\right)$ be the proportion of

cases *in only the discordant cells* (DD for both discordant cells) where only the tester in the control condition (c) receives an affirmative response, and $p_{DD} = \left(\frac{n_{10}+n_{01}}{n}\right)$ be the proportion of cases in both discordant cells relative to the total. The formula for the sample size of McNemar's test with a set significance level α and power of $1 - \beta$ is given by:

$$n = \frac{\left(z_{1-\alpha/2} + 2z_{1-\beta}\sqrt{p_{c/DD}\left(1 - p_{c/DD}\right)}\right)^2}{4\left(p_{c/DD} - .5\right)^2 p_{DD}}, \tag{3}$$

and the formula for the power for McNemar test with a set α, sample size n, and Φ as the cumulative distribution function of the standard normal distribution is as follows:

$$1 - \beta = \Phi\left[\frac{1}{2\sqrt{p_{c/DD}\left(1 - p_{c/DD}\right)}}\left(z_{\alpha/2} + 2\left|p_{c/DD} - .5\right|\sqrt{np_{DD}}\right)\right]. \tag{4}$$

The results are symmetric, such that the same sample size and power emerge when the values of π_{10} and π_{01} are interchanged. For a one-sided hypothesis, $\alpha/2$ is replaced by α.

All calculations were obtained with the statistical software program R (R Development Core Team 2006). We created functions for calculating power and sample size for McNemar's test. Although some functions are available in the "TrialSize" package, we created a function with additional features. For example, our function allows for both one-sided and two-sided hypotheses. We also created a function that automatically produces a sample size table for a given α and β (as equivalent to Table 4) and power calculations for a given sample size (as equivalent to Table 5). In addition, functions are available for the sample size for the Stuart–Maxwell test and Cochran's Q test. For the latter, this function represents the first of its kind, which is important given the increasing use of more than one treatment in audit studies. These functions are available on the first author's website.

Sample Size Calculations, Choices, and Challenges

Table 4 shows the sample sizes that result from the one-sided versions of the formulas above with $\alpha = .05$ and $1 - \beta = .80$. For McNemar's test, we present the one-sided case because it is directly applicable to our methods in the empirical example that follows, though the formulas, functions, and

Table 4. Sample Size Required for a Power of 0.80 for McNemar's Test.

		p_{01} = Proportion of *units* where treatment tester *in pair* does receive an affirmative response when control tester *in pair* does not								
		0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
p_{10} = proportion of *units* where control tester *in pair* does receive an affirmative response when treatment tester *in pair* does not	0.10	357								
	0.15	113	610							
	0.20	60	179	860						
	0.25	39	90	243	1,109					
	0.30	28	57	119	305	1,347				
	0.35	22	40	73	148	368	1,605			
	0.40	18	30	51	90	176	430	1,852		
	0.45	15	24	38	61	106	204	492	2,100	
	0.50	13	20	30	45	72	122	232	555	2,348
	0.55	11	17	24	35	53	82	137	259	617
	0.60	10	14	20	29	41	60	92	153	
	0.65	9	13	17	24	33	46	67		
	0.70	8	11	15	20	27	37			
	0.75	7	10	13	18	23				
	0.80	7	9	12	15					
	0.85	6	8	11						
	0.90	6	8							
	0.95	5								

Note: $\alpha = .05$ and $1 - \beta = .80$. Sample size is for the *experimental unit*. Shaded area represents sample sizes higher than 300, as used in our empirical example. Highlighted box represents the assumed likely distributions in the empirical example.

calculations are easily extended and analogous to the two-sided case.[5] In what follows, we adopt terminology whereby the one-sided hypothesis favors higher values in the cell in which the control receives affirmative responses and the treatment does not (H_0: $\pi_{10} > \pi_{01}$), though the symmetric nature of the calculations allows for the reverse to be considered by simply swapping the notation between treatment and control. In employment discrimination studies, higher values are often anticipated in the cell in which only the control receives an affirmative response. As noted, the results depend on the percentage in each of the discordant cells, though the total concordant cells (regardless of the distribution across the two concordant cells) will be determined by the proportions that fall into the discordant cells. Thus, the rows of Table 4 represent the proportion of all experimental units where the control tester received an affirmative response when the treatment tester

Table 5. Power for McNemar's Test with Sample Size of 300.

		p_{01} = Proportion of *units* where treatment tester *in pair* does receive an affirmative response when control tester *in pair* does not								
		0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
p_{10} = Proportion of *units* where control tester *in pair* does receive an affirmative response when treatment tester *in pair* does not	0.10	0.74								
	0.15	0.99	0.54							
	0.20	1.00	0.95	0.43						
	0.25	1.00	1.00	0.87	0.36					
	0.30	1.00	1.00	0.99	0.79	0.32				
	0.35	1.00	1.00	1.00	0.97	0.73	0.28			
	0.40	1.00	1.00	1.00	1.00	0.95	0.67	0.26		
	0.45	1.00	1.00	1.00	1.00	1.00	0.92	0.62	0.24	
	0.50	1.00	1.00	1.00	1.00	1.00	0.99	0.88	0.57	0.22
	0.55	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.85	0.53
	0.60	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	
	0.65	1.00	1.00	1.00	1.00	1.00	1.00	1.00		
	0.70	1.00	1.00	1.00	1.00	1.00	1.00			
	0.75	1.00	1.00	1.00	1.00	1.00				
	0.80	1.00	1.00	1.00	1.00					
	0.85	1.00	1.00	1.00						
	0.90	1.00	1.00							
	0.95	1.00								

Note: $\alpha = .05$. Shaded area represents power below 0.80. Highlighted box represents the assumed likely distributions in the empirical example.

did not, or $p_{10} = \left(\frac{n_{10}}{n}\right)$. The columns represent the proportion of all experimental units where the treatment tester received an affirmative response when the control tester did not, or $p_{01} = \left(\frac{n_{01}}{n}\right)$. We can then represent the total proportion in both concordant cells as $p_{CC} = 1 - p_{01} - p_{10}$.

By examining Table 4, we can compare required sample sizes for similar magnitude differences. For example, the very first entry in Table 4 where $p_{10} = 0.10$ and $p_{01} = 0.05$ has the lowest value of p_{DD}, with $p_{DD} = 0.05 + 0.10 = 0.15$ and $p_{CC} = 1 - 0.15 = 0.85$. In the design phase, there is thus an assumption that for this five percentage point difference in the outcome, 85 percent of the cases fell into the concordant

cells where neither or both tester received a callback. Only the control received an affirmative response 10 percent of the time and only the treatment received an affirmative response 5 percent of the time. The necessary sample size to detect this difference is 357. The cell for $p_{10} = 0.15$ and $p_{01} = 0.10$ also has a five percentage point difference in the outcome, but here $p_{CC} = 0.75$ and the necessary sample size to detect *this* five percentage point difference is much larger at 610. Similarly, all cells following that first diagonal exhibit a five percentage point difference, but require very different sample sizes to maximize the chances to detect *the same size effect*. As shown in the table, power is lower as p_{10} and p_{01} *simultaneously* approach 0.5 (or that the sum equals 1). For example, consider two extreme cases. First, in the cell where $p_{10} = 0.55$ and $p_{01} = 0.45$, we have $p_{CC} = 0$ (i.e., all the pairs were discordant) and the necessary sample size is 617. Second, in the cell where $p_{10} = 0.75$ and $p_{01} = 0.25$, we still have $p_{CC} = 0$, but the necessary sample size is only 23 given the very large magnitude difference.[6]

We use these examples to demonstrate the difficulty of considering only the magnitude difference when planning a matched field experiment with a dichotomous outcome. Instead, the proportion in the concordant cells and the breakdown in the respective discordant cells influence sample size choice. Researchers' expectations of these quantities, together with the hypothesized effect size they would like to detect, must guide sample size selection. As will follow in our example and recommendations, we suggest that when designing a paired experiment with a dichotomous outcome, it is best to identify the realistic areas of Table 4 and choose a sample size that maximizes the chances of finding a significant effect for a given sample size across several possible outcome distributions. Both the concordant proportion and the amount in each discordant proportion are difficult to determine a priori. One would need information about how often either both or neither of the testers receives an affirmative response, and, for the discordant cells, how often the hypothetically unlikely case occurs when the treatment tester receives an affirmative response while the control tester does not. This challenge is best illustrated with an empirical example, to which we now turn.

Application to an Empirical Example

Despite the difficulties inherent in predicting concordance versus discordance and the separate discordant proportions in advance of a study, we can still calculate power for various sample sizes and show how this varies across

these factors. We demonstrate this exercise with an empirical example (described in full in Uggen et al. 2014). In an audit study modeled after Pager (2003), we sent same-race pairs to 300 randomly selected establishments in the Twin Cities metropolitan area, with one applicant reporting no criminal history and the other reporting a misdemeanor arrest record. Four young male testers applied for entry-level jobs using fictitious identities.[7] All entry-level advertisements were selected, so long as they required no special skills or licenses, instructed applicants to apply in-person, and were located in the seven-county Twin Cities metropolitan area. Each week from August 2007 to June 2008, one tester in each pair was assigned to the treatment condition, that is, a single misdemeanor disorderly conduct arrest. Over the course of eight months, each pair submitted close to 300 applications at 150 job sites, with each tester assigned to the treatment condition for half of the audits. Our primary dependent variable was an employer "callback," measured by an offer of employment or an invitation for a second interview (whether in-person or through e-mail or voicemail).

In selecting our sample size, we chose such that a reasonable difference could be detected across most of the distribution where realistic values of p_{10} and p_{01} fell, while staying within the confines of the available budget. This reasonable difference is typically between 5 to 10 percentage points. Here, we describe the calculations that led to this choice.

As described in the previous section, in the design phase of a matched field experiment, researchers must determine two numbers that are difficult to know a priori: the expected proportions in both discordant cells, as determined by the difference one wishes to detect, and the expected amount in the concordant cells. We viewed high values of p_{01} as unlikely. That is, we assumed it would be rare for the otherwise equally qualified applicant with the arrest record to receive a callback when the applicant with a clean record did not. (Rather, we viewed concordance as *much* more likely: If the treatment tester received a callback, then the control was expected to as well). In one respect then, we used such likelihoods as anchors to determine the section of Table 4 into which our expectations fell. For the other unknown, we also viewed a low amount of total concordance as unlikely. That is, we assumed that there would be many employers who would call back both or neither tester (although irrelevant for power, we particularly viewed the latter as constituting many of the employers). Thus, we saw the outlined box in Table 4 as the area in which our study design was realistically positioned. Within this box, p_{10} reaches a maximum of 0.60, while p_{01} does not exceed 0.15. Depending on the distribution between concordant and discordant as well as the two discordant cells' distance to 0.5, we have the power to detect

Table 6. Distribution of Callbacks by Criminal Record for Each Paired Audit and McNemar's Test.

		Misdemeanor Arrest		
		Callback	No Callback	Total
No Misdemeanor Arrest	Callback	$n_{11} = 60$ $p_{11} = .20$	$n_{10} = 39$ $p_{10} = .13$	$n_{1+} = 99$ $p_{1+} = .33$
	No Callback	$n_{01} = 27$ $p_{01} = .09$	$n_{00} = 174$ $p_{00} = .58$	$n_{0+} = 201$ $p_{0+} = .67$
	Total	$n_{+1} = 87$ $p_{+1} = .29$	$n_{+0} = 213$ $p_{+0} = .71$	$n = 300$ $p_{++} = 1$

Note: $\chi^2 = 1.833$, $df = 1$, $p = .088$ (one-sided hypothesis); Odds ratio $= 1.44$.

an effect should it exist in the population at a magnitude difference of between 5 and 10 percentage points.

Put another way, within the confines of the resources available to us, we were comfortable with a low power to detect an effect where the percentage point difference *in the discordant cells* was only 5 to 10 percentage points. Once a sample size is selected, we can reverse the question and consider the exact magnitude difference that is detectable for different levels of concordance. As Table 5 demonstrates, considering the power across all the possible concordant and discordant combinations while restricting the sample size to 300, power is quite high when the magnitude difference exceeds five percentage points in our highlighted box. For example, with 300 employers and an assumption that $p_{01} = 0.05$ (the first column of the highlighted box), the threshold where power 0.80 is crossed equates to a 5.55 percentage point difference (i.e., $p_{10} = 0.1055$). If instead we are located in the last column of our highlighted box where $p_{01} = 0.15$, we can detect an 8.9 percentage point difference (i.e., $p_{10} = 0.2390$) with a power of 0.80.

We now discuss the observed results from our audit study (for a full discussion, see Uggen et al. 2014). Table 6 displays the distribution of callbacks. According to the marginal callback rate, testers with no misdemeanor arrest received a callback to 33.0 percent of their applications, while those with an arrest received a callback to 29.0 percent of their applications. These marginal proportions, however, do not take into account the result of the other tester in the pair at the same employer, which would lead to potentially erroneous conclusions if a χ^2 test was conducted on this cross tabulation or a z-test of proportions (which assume independence).[8] In fact, Donner and Li (1990) showed that McNemar's test is a function of a Pearson χ^2 and a

kappa statistic. The latter can be interpreted as the estimated correlation between members of a matched pair under the null, such that the relative magnitudes of the classical unmatched and matched χ^2 statistics depend solely on the degree of resemblance within members of a matched pair (Donner and Li 1990:828; cf. Newcombe 1996).

Unfortunately, articles do not consistently report the *pair-specific* cross tabulations, which is the information future researchers need to use past studies as guides to calculating sample size and power. The marginal callback rate that is often given in matched audit studies prior to statistical modeling does not provide the needed pair-specific outcomes, which are shown in the pair-specific cross tabulation in Table 6. For example, while 33.0 percent of testers without an arrest and 29.0 percent of testers with an arrest received a callback, only 13.0 percent of the former received a callback when the latter did not and 9.0 percent of the latter received a callback when the former did not. If future researchers want to use past research as a starting point for sample size and power calculations, these pair-specific numbers are necessary information. Comparing our observed results with the expectations in our sample size analysis also reveals some of the difficulties in calculations for matched designs.

Given the *observed* concordance and relatively low percentage point difference in the discordant cells in our conducted study, the power to detect such a difference is relatively low, consistent with our a priori assumptions.[9] Compared to our assumptions, we observed a much higher proportion of concordance (p_{CC}) and proportion of employers where only the tester *with* the record received a callback (p_{01}) relative to the proportion of employers where only the tester *without* the record received a callback (p_{10}). In other words, we assumed that for such a high callback rate for the treatment tester only, the control tester only would receive a higher number of callbacks than was observed. Across the sample of 300 employers, $p_{CC} = 0.78$, $p_{01} = 0.09$, and $p_{10} = 0.13$. Each of these numbers implies that we are near the very top of our outlined boxes in our tables. That is, even though they were otherwise comparable on all characteristics (and with the misdemeanor record rotated between the testers), the tester with the record received a callback when the tester without the record did not at a full 9 percent of the employers. Thus, we encourage researchers to prepare for this nonintuitive cell in their sample size calculations, which we clearly underestimated relative to p_{01}, as well as to collect covariates that might explain this anomaly, points to which we return in our recommendations.

Another important consideration is the degree of concordance. In addition to the observed discordance, the concordance of 78 percent was much higher

than anticipated. The concordant cells are often difficult to determine beforehand. Although only their sum is relevant, understanding the context of each cell is necessary to make a realistic conclusion about the sum. For example, the 58 percent of the employers who called neither tester was highly dependent on local labor market conditions. Thus, even with information from a prior study, this cell, and thus the sum of the concordant cells, may be difficult to assign. In the cell in which both testers received a callback, constituting 20 percent of the sample, the amount is dependent on the relative "strength" of the treatment. Relatively "weak" treatments (here a misdemeanor arrest as opposed to Pager's 2003 felony prison record) are correspondingly more likely to be overlooked. In such cases, we would expect high concordance in the affirmative cell, but would still have little reason to expect only the tester presenting the record to receive a callback.

As is typical for power curves, the relationship is not linear. As Table 4 shows, the sample size shrinks precipitously, as the magnitude of the percentage point difference in the two discordant cells increases and as this same difference extends farther from 0.5 (i.e., higher concordant percentages). Although power is most appropriately considered a design question, this result implies that even small changes in our observed data between these cells closer to our assumptions would have resulted in an acceptable power level. For example, if just 10 of the 60 employers who called back both testers instead called only the tester without the record, the power to detect the resultant difference would be 0.82. Similarly, if just 7 of the 27 employers in our unlikely cell who called back only the tester with the record had instead called back neither tester, the power to detect this difference would be 0.81. Thus, even though the power to detect the observed difference is low, small shifts in our observed data toward our a priori assumptions would have made the power more reasonable. This result emphasizes the need to consider a large range of possible outcome distributions when selecting sample sizes.

Stuart–Maxwell Test: More Than Two Outcome Categories

Although not yet considered in the social science literature, we could also easily envision scenarios in which there are still two paired testers, but the outcome has more than two response categories. This might arise, for example, if African American applicants are "channeled" to a lower position than the one advertised, while Whites are not (see, e.g., Pager, Western, and Bonikowski 2009). In such a scenario, both applicants would receive a positive response from employers, but we might want to know whether the type of

response differs by race. Thus, there would be two testers but three outcomes: (1) callback for advertised position or higher, (2) callback but channeled downward, and (3) no callback. Not surprisingly, the challenges of sample size selection become greater as the number of outcome categories grows. Therefore, it is important to also consider sample size and power calculations for the case of these contingency tables.

As noted, the case of McNemar's test generalized to more than two response categories is known as a Stuart–Maxwell test (Stuart 1955; see also Agresti 2002:458–59).[10] This test still assesses the assumption of marginal homogeneity, H_0: $\pi_{i+} = \pi_{+i}$, but instead does so simultaneously for all i, with $i = 1, 2, \ldots, r$ response categories. This again amounts to simultaneously testing the equality of all possible discordant combinations, or H_0: $\pi_{ij} = \pi_{ji}$ for all $i, j = 1, \ldots, r$, $i \neq j$. Thus, r response categories and $r(r-1)/2$ null hypotheses are simultaneously tested. More intuitively, this hypothesis tests whether the proportion for each possible combination of outcomes is the same for the two treatments. Table 7 shows notation analogous to that presented for McNemar's test for the Stuart–Maxwell test for $r = 3$ outcome categories, with an accompanying example following that given earlier. For r response categories, the Stuart–Maxwell test statistic is:

$$\chi^2_{r(r-1)/2} = \sum_{i<j} \frac{\left(n_{ij} - n_{ji}\right)^2}{n_{ij} + n_{ji}}. \tag{5}$$

Based on the derivation of the exact test, Chow, Shao, and Wang (2003:154–55) present the following sample size formula:

$$n = \lambda_{\alpha,\beta} \left[\sum_{i<j} \frac{\left(p_{ij} - p_{ji}\right)^2}{p_{ij} + p_{ji}} \right]^{-1}, \tag{6}$$

where $\lambda_{\alpha,\beta}$ is the noncentrality parameter from a noncentral χ^2 distribution for a desired α and β and r degrees of freedom.

As with the case of two outcome categories, the determination of sample size, as well as the test itself, only uses information from the discordant cells. Again, this calculation amounts to having an estimate of the total proportion in the concordant cells and each of the discordant cells beforehand. We briefly demonstrate sample size calculations for the case of $r = 3$. Here, the $r(r-1)/2 = 3$ null hypotheses being simultaneously tested are $\pi_{12} = \pi_{21}$, $\pi_{13} = \pi_{31}$, and $\pi_{23} = \pi_{32}$. For the above-mentioned example, the null hypotheses would state that African Americans and Whites receive the same percentage for all three possible discordant comparisons. As in analysis of variance

Table 7. Notation for Stuart–Maxwell Test for Case of Three Response Categories (with Hypothetical Example for Racial Discrimination in Hiring).

		Treatment (Black)			
		Outcome 1 (Callback, Advertised Position or Higher)	Outcome 2 (Callback, Lower Position than Advertised)	Outcome 3 (No Callback)	Total
Control (White)	Outcome 1 (Callback, advertised position/higher)	Population proportion: π_{11} Sample proportion: p_{11} Sample cell size: n_{11}	Population proportion: π_{12} Sample proportion: p_{12} Sample cell size: n_{12}	Population proportion: π_{13} Sample proportion: p_{13} Sample cell size: n_{13}	Population proportion: π_{1+} Sample proportion: p_{1+} Sample cell size: n_{1+}
	Outcome 2 (Callback, lower position)	Population proportion: π_{21} Sample proportion: p_{21} Sample cell size: n_{21}	Population proportion: π_{22} Sample proportion: p_{22} Sample cell size: n_{22}	Population proportion: π_{23} Sample proportion: p_{23} Sample cell size: n_{23}	Population proportion: π_{2+} Sample proportion: p_{2+} Sample cell size: n_{2+}
	Outcome 3 (No callback)	Population proportion: π_{31} Sample proportion: p_{31} Sample cell size: n_{31}	Population proportion: π_{32} Sample proportion: p_{32} Sample cell size: n_{32}	Population proportion: π_{33} Sample proportion: p_{33} Sample cell size: n_{33}	Population proportion: π_{3+} Sample proportion: p_{3+} Sample cell size: n_{3+}
	Total	Population proportion: π_{+1} Sample proportion: p_{+1} Sample cell size: n_{+1}	Population proportion: π_{+2} Sample proportion: p_{+2} Sample cell size: n_{+2}	Population proportion: π_{+3} Sample proportion: p_{+3} Sample cell size: n_{+3}	Population proportion: $\pi_{++} = 1$ Sample proportion: $p_{++} = 1$ Sample cell size: n

(ANOVA), we reject the null for the overall Stuart–Maxwell test when there is evidence against any of these null hypotheses. In the case of two outcomes, the required sample size grew as the difference between the discordant proportions decreased and as the sum of the discordant proportions simultaneously grew closer to 1. For two outcomes, this meant the discordant proportions simultaneously approached 0.5. For r outcomes, this implies that the discordant proportions simultaneously approach an even split, or $1/[r\,(r-1)]$. So for the case of three outcome categories, the highest sample size is required when the proportion in each discordant cell is close to 0.167. As the gap between the comparisons grows, a lower sample size is required. But this is offset by the total amount in the concordant cells. As is the case with two outcomes, the same two factors again determine sample size: the difference between the discordant cells and the total amount in the concordant cells. The difference here, of course, is in the former, where any given comparison can alter the sample size calculation.

For example, if we want to detect the same five percentage point difference in all the outcomes between the control and the treatment and assume in our distribution that all $\pi_{ij} = .13$ and all $\pi_{ji} = .18$ (and thus a total of 0.07 in the concordant cells), then we would need a sample size of 451 for $\alpha = .05$ and $1 - \beta = .80$. If, however, the five percentage point difference assumes all $\pi_{ij} = .05$ and $\pi_{ji} = .10$ (and thus a total of 0.55 in the concordant cells), the required sample size would be 219. Again, the required sample size to achieve the same power is not consistent for the same percentage point difference. In these examples, we assume the proportion in each of the discordant outcomes for the treatment and control are the same, respectively, but there is no reason to assume such.

To depict the required sample sizes for our example of three response categories, we chose several illustrative comparisons and graphed them in Figure 1. We take a graphical approach because reproducing a table similar to Table 4 for McNemar's test is challenging since we can only show one π_{ij} versus π_{ji} comparison, while fixing the others at specific values. In other words, the sample size values in the table for one π_{ij} versus π_{ji} comparison differ depending on the proportions in the remaining discordant cell comparisons. In the figure, we concentrate on the π_{12} versus π_{21} comparison. In our example, these values represent the proportion of employers that call back the White or the Black tester, respectively, for the job advertised or higher (outcome 1), while the other tester receives a callback for a lower position (outcome 2). We vary the assumed proportion for the control group (π_{12}, or the White tester receives the more favorable response) in the figure, while fixing its comparison to the treatment group (π_{21}, or the Black tester receives the more favorable response) at 0.05, 0.15, and 0.25.

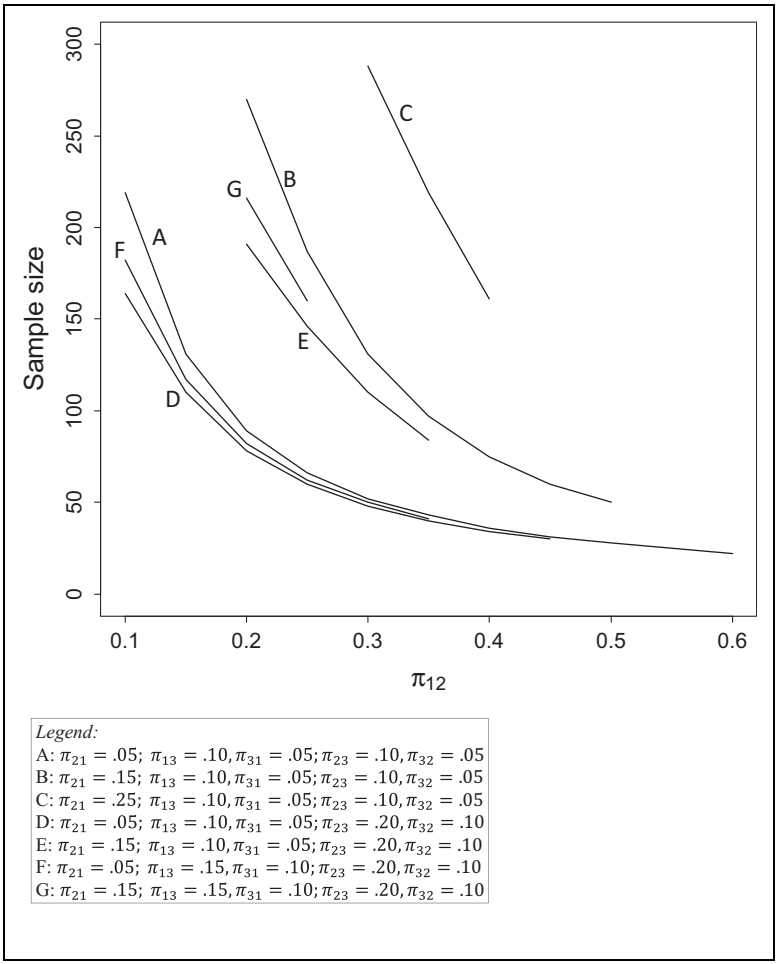

Figure 1. Required sample size for Stuart–Maxwell test. *Note.* $\alpha = .05$ and $1 - \beta = .80$.

Lines A, B, and C fix the other two outcome comparisons at 0.05 and 0.10 for the treatment and control, respectively. Reading left to right from the y-axis, each line begins with a five percentage point difference, yet the sample sizes differ because the total amount in the concordant cells decreases as π_{21} increases. At the beginning of line A, $\pi_{12} = .10$ and $\pi_{21} = .05$, replicating the previous example with a required sample size for power 0.8 of 219. At the

beginning of line B, $\pi_{12} = .15$ and $\pi_{21} = .20$. With this decrease in the total concordant cells, the required sample size for this five percentage point comparison is 270. If we go further with line C where $\pi_{12} = .25$ and $\pi_{21} = .30$, the required sample size is 288. For each line, as we only vary π_{12}, the required sample size decreases as the percentage point difference for the comparison under consideration increases across the x-axis.

The remaining lines then vary the other two comparisons. (Since the proportions must add to 1, the lines are of varying lengths.) The second comparison (π_{13} vs. π_{31}) represents the proportion of employers where the White or Black tester, respectively, receives a callback for the advertised position or higher (outcome 1), while the other tester receives no callback (outcome 3). We keep the same five percentage point difference for lines D and E, while shifting to .15 and .10 for lines F and G. The third comparison (π_{23} vs. π_{32}) represents the proportion of employers where the White or Black tester, respectively, receives a callback for a lower position than advertised (outcome 2), while the other tester receives no callback (outcome 3). Across lines D through G, there is a 10 percentage point difference (.20 vs. .10) for this comparison. With the larger difference for that comparison, these four lines are lower than lines A through C. For lines F and G, the second comparison is still a five percentage point difference, but results in a lower proportion in the concordant cells, and thus requires a higher sample size compared to lines D and E. The figure thus illustrates how the two challenges associated with a priori assumptions when calculating sample size also apply to more than two outcomes (i.e., the total amount in the concordant cells and the relative proportions in the discordant cells). Of course, the increased difficulty comes in assigning proportions to *three* discordant cell pairs. Although it is nontrivial to add an additional outcome, researchers may have good theoretical reasons to do so, such as the hypothetical example provided previously. Although there is increased difficulty in determining sample size, there may be relatively little cost, yet great scientific yield, associated with increasing the number of outcomes. By using the functions we provide, however, some of the difficulty in sample size determination is alleviated. We return to these points in our recommendations.

Cochran's Q Test: More Than Two Repeated Measures

Although several of the audit studies reviewed earlier only use two testers per experimental unit, some have used three or more repeated observations. For example, Pager, Western, and Bonikowski (2009) sent White, Latino, and

African American applicants to the same employer, and Bertrand and Mullainathan (2004) sent four applications to the same employer for each combination of White or African American of high and low qualifications. Sample size calculations for extending to more than two treatments are less straightforward than extending to more than two outcome categories. As mentioned previously, the case of McNemar's test generalized to more than two repeated observations is known as Cochran's Q statistic (Cochran 1950; see also Agresti 2002:458–59; Patil 1975; Wallenstein and Berger 1981). This test also assesses a version of marginal homogeneity, H_0: $\pi_{1++\ldots+} = \pi_{+1+\ldots+} = \ldots = \pi_{+++\ldots1}$ for the m treatments (which is the number of subscripts). More intuitively, it tests whether the treatments have the same effect—whether the difference in the proportion of affirmative responses to each treatment is zero in the population. Note the differences between the case of r outcomes and m treatments. For the r outcomes, the Stuart–Maxwell test assesses whether a series of bivariate comparisons in an $r \times r$ table are all equal. For the m outcomes, Cochran's Q simultaneously tests whether all the treatments in m 2×2 tables are equal.

We can again rewrite the hypotheses in terms of just the discordant cells. We use the example of three treatments to illustrate. Unlike the case of additional outcome categories where we can visualize the proportions in an $r \times r$ table (see Table 7), when we add additional treatment categories, we instead increase the dimensions of the table. In a visual format suggested in Cochran's (1950) original article, Table 8 shows the notation for three treatments. Thus, the null hypothesis is H_0: $\pi_{1++} = \pi_{+1+} = \pi_{++1}$. When we rewrite the null as a series of differences, we see again that the test only depends on discordant cells:

$$\begin{aligned}\pi_{1++} - \pi_{+1+} &= (\pi_{111} + \pi_{110} + \pi_{101} + \pi_{100}) - (\pi_{111} + \pi_{110} + \pi_{011} + \pi_{010}) \\ &= (\pi_{101} + \pi_{100}) - (\pi_{011} + \pi_{010}).\end{aligned} \tag{7}$$

Similarly,

$$\pi_{1++} - \pi_{++1} = (\pi_{110} + \pi_{100}) - (\pi_{011} + \pi_{001}), \text{and,} \tag{8}$$

$$\pi_{+1+} - \pi_{++1} = (\pi_{110} + \pi_{010}) - (\pi_{101} + \pi_{001}). \tag{9}$$

Essentially, each difference removes the common occurrence of an affirmative response from the comparison, since the treatments are equally effective in that case.

Table 8. Table for Cochran's *Q* Test for Three Matched Proportions.

	Treatment 1 Response	Treatment 2 Response	Treatment 3 Response	Population Proportion	Sample Proportion	Sample Cell Size
	1	1	1	π_{111}	p_{111}	n_{111}
	1	1	0	π_{110}	p_{110}	n_{110}
	1	0	1	π_{101}	p_{101}	n_{101}
	0	1	1	π_{011}	p_{011}	n_{011}
	1	0	0	π_{100}	p_{100}	n_{100}
	0	1	0	π_{010}	p_{010}	n_{010}
	0	0	1	π_{001}	p_{001}	n_{001}
	0	0	0	π_{000}	p_{000}	n_{000}
Marginal population proportion	π_{1++}	π_{+1+}	π_{++1}			
Marginal sample proportion	p_{1++}	p_{+1+}	p_{++1}			
Marginal sample size	n_1	n_2	n_3			

For m treatments and A_i representing the number of affirmative responses by experimental unit i, $i = 1, \ldots, n$, such that $A_i \in \{0, 1, 2, \ldots, m\}$, we have:

$$Q = \frac{m(m-1)\sum_{j=1}^{m} n_j^2 - (m-1)\left(\sum_{j=1}^{m} n_j\right)^2}{m\sum_{j=1}^{m} n_j - \sum_{i=1}^{n} A_i^2}, \tag{10}$$

where Q is distributed as χ^2_{m-1} (Cochran 1950). For the statistic, the A_i term implies that one also needs information from the case in which all the responses are affirmative, rather than strictly the discordant cells as in the case of two treatments. The same is also true of sample size calculations.

Calculating sample size for Cochran's Q requires a unique approach. To our knowledge, no one has attempted to explicitly state an exact formula or create a function to calculate sample size for Cochran's Q, due to the difficulty in inverting the formula under the alternative hypothesis to solve for N. In light of the relationship between Pearson's χ^2 statistic and Cochran's Q test, Donner and Li (1990) state that sample size calculations can be computed by multiplying the required sample size for the independent contingency table for each treatment (see Lachin 1977) by a kappa statistic κ_m for intraclass correlation (Cohen 1960; Fleiss 1981; Scott 1955). Thus, we provide a function that takes the sample size calculation for an independent $m \times 2$ table of the marginal proportions as given in Lachin (1977), which we denote $n_\perp$, and multiplies this value by a reexpressed version of the weight proposed by Donner and Li (1990:831), which we call w_Q:

$$n = w_Q n_\perp, \tag{11}$$

where

$$w_Q = 1 - \kappa_m = \frac{\sum_{j=0}^{m} j(m-j)p_j}{m(m-1)\sum_{j=0}^{m} \frac{jp_j}{m}\left(1 - \sum_{j=0}^{m} \frac{jp_j}{m}\right)}, \tag{12}$$

and p_j represent the proportion of units with affirmative response value j, $j \in \{0, 1, 2, \ldots, m\}$. For more detail, Online Appendix A shows the derivation of $n_\perp$ from Lachin (1977) as it applies to Cochran's Q, and Online Appendix B demonstrates the algebraic equivalency of $1 - \kappa_m$ from Donner and Li (1990) and w_Q.

With regard to the weight, a kappa statistic measures interrater reliability. When it is less than zero, there is poor agreement. When it is between 0 and 1, it measures the degree of similarity. Thus, $w_Q = 1 - \kappa_m$ assigns higher weights, and thus sample size, when the agreement is poor, which here

implies higher discordance. The weight is smaller when the concordance is high. With more than two treatments, however, there are different degrees of concordance based on the number of affirmative responses within a given experimental unit (e.g., π_{111} is "more concordant" than π_{110}, although the latter is "equally concordant" to π_{100} because the number of like responses is the same). Through the lead coefficient in the numerator, w_Q thus weights more heavily those cells that are "more discordant," which becomes especially relevant as the number of treatments grows. When the responses are "completely" concordant (i.e., either all or no affirmative responses), it is easy to see that those cases do not contribute to the numerator, as the lead coefficient is zero. Like the test statistic itself, however, knowledge of the proportion of experimental units with all affirmative responses is needed to calculate the denominator. Although technically one does not need the proportion for the cell where no testers receive an affirmative response, this cell will be predetermined since all other cells are required.

As with the previous two statistical tests, the weight results in a higher required sample size as both the total concordance decreases and the differences in the marginal proportions decrease. Demonstrating this numerically is more difficult than with either McNemar's test or the Stuart–Maxwell test because there is no direct comparison between two proportions. As the null hypotheses show, each marginal comparison actually contains several discordant proportions. Further, each null contains *overlapping* proportions for cells where more than one affirmative response occurred. Unfortunately then, altering a given proportion can affect more than one null comparison, increasing the difficulty in considering a range of possible sample sizes. In fact, the researcher must provide *all* the proportions. In the case of three treatments, there are eight proportions, as demonstrated in Table 8. For four treatments, there are 16 proportions, which is shown in Online Appendix C. Thus, considering values required for sample size calculations quickly becomes onerous as m increases.

We provide a format for researchers calculating sample size for Cochran's Q in Table 9. In this hypothetical example, we consider an audit of racial discrimination in hiring using White, Black, and Latino applicants. The statistical test is evaluating whether the marginal proportions are equivalent, given the clustering. In example 1, Whites received a callback in 44 percent of applications, Latinos in 23 percent of applications, and Blacks in 16 percent of applications. Simply computing a sample size for these numbers as independent, however, would ignore that each margin shares constituent proportions with the other treatments. Thus, above the marginal values, we provide the matched proportions that led to these marginal proportions, in which

Table 9. Cochran's *Q* Sample Size for Hypothetical Test of Racial Discrimination in Hiring.

	White Response	Latino Response	Black Response	Sample Proportion	Empirical Callback Results	Example Proportions 1	Example Proportions 2	Example Proportions 3
	1	1	1	p_{111}	All	0.05	0.08	0.05
	1	1	0	p_{110}	White and Latino	0.13	0.16	0.12
	1	0	1	p_{101}	White and Black	0.08	0.10	0.08
	0	1	1	p_{011}	Latino and Black	0.02	0.05	0.06
	1	0	0	p_{100}	White only	0.18	0.21	0.12
	0	1	0	p_{010}	Latino only	0.03	0.06	0.04
	0	0	1	p_{001}	Black only	0.01	0.04	0.03
	0	0	0	p_{000}	None	0.50	0.30	0.50
Marginals	π_{1++}	π_{+1+}	π_{++1}	p_{1++}	White rate	0.44	0.55	0.37
	p_{1++}	p_{+1+}	p_{++1}	p_{+1+}	Latino rate	0.23	0.35	0.27
	n_1	n_2	n_3	p_{++1}	Black rate	0.16	0.27	0.22
					Sample size	103	145	372

Note: $\alpha = .05$ and $1 - \beta = .80$.

Whites are assumed to receive more callbacks, while Latinos and Blacks (either solely or together) receive fewer callbacks, and the proportion of concordance is 0.50. In this example, the required sample size is 103 for $\alpha = .50$ and $1 - \beta = .80$. In example 2, we consider a case with less of the overall proportion in the concordant cells at 0.30. Yet, we distribute that 0.20 decrease such that the marginals remain of the same distance from one another (within one percentage point due to rounding). Here, decreasing the amount of concordance results in a required sample size of 145. Thus, for an equivalent percentage point difference in the marginals, we again require differing sample sizes. In example 3, we keep the proportion of concordance the same as in example 1, but reduce the level of discrimination. Even with what amounts to very small shifts in the various discordant proportions, the required sample size increases to 372. Thus, shifts in both concordance and discordance again contribute to sample size, but here they are complicated by the fact that each marginal contains multiple constituent proportions.

Cautions for Extensions of the Method

In the social sciences, researchers also typically display linear models. For continuous outcomes, ANOVA and linear regression with only the treatment effect(s) in the model lead to the same results for power, sample size, and inferential conclusions. When data are clustered, fixed effects (FE) logit models with only the treatment(s) effects are analogous and lead to similar conclusions, although differing estimation procedures may produce slight differences in the bivariate test. In the case of equal exposure within experimental units to both the treatment and the control, as for audit studies, the bivariate tests are also equivalent to a random effects (RE) logit model with just the treatment effect as a predictor. Given this equivalency, the choice of RE or FE for the bivariate logit model is of little consequence, and the calculations here are easily extended to the generalized linear model with only the treatment effect included.

Typically, many articles employ an additional explanatory variable of interest (e.g., race in Pager 2003; gender in Correll et al. 2007). These variables, however, do not conform to the typical definition of a stratification variable within the generalized CMH framework, in which the experimental unit is contained within a mutually exclusive category of a confounding variable that cannot be randomly assigned (e.g., employers *are not contained within* race in Pager 2003 or *within* gender in Correll et al. 2007; cf. employers *contained within* states in Tilcsik 2011). Those variables also cannot be considered another treatment level because it was not varied within

experimental units (cf. Bertrand and Mullainathan 2004; Pager, Western, and Bonikowski 2009). Thus, these examples are more akin to separate McNemar's tests by the second variable of interest. By design, they cannot covary with the treatment, as no variation exists within employers on that measure (essentially constituting a covariate rather than a treatment, which is why we consider them as such in this discussion). Similarly, covariates collected during an audit would not constitute stratifying variables because they cannot be known beforehand. Nevertheless, these cases highlight the importance and utility of additional statistical modeling, which we consider further in the discussion section. We note here that, in such cases where additional covariates are included, RE logit models are likely the preferred choice, as several important covariates, including those listed in the studies previously, only vary across experimental units and thus are not estimable in the FE framework.

Although we might be interested in an additional stratifying variable as opposed to a covariate, such a test, along with its corresponding sample size formula, is not developed in the literature specifically for *matched* pairs. This would constitute a four-way table with O outcome response categories, R repeated measures, $S = N$ strata for each experimental unit, and some other stratification variable to which each experimental unit would a priori belong (e.g., Tilcsik's [2011] interest in the effect of state in an audit of employers' hiring of gay men might conform to a four-way design). For multidimensional contingency tables, researchers usually turn to log-linear models for cell counts. For matched designs, however, this is not possible because the strata variable with N categories would have to be included, which would imply that every cell only contains either a 1 or a 0 (see Agresti 2002:233–34, 413–14, table 10.2), which would create problems in their estimation (Agresti 2002:391–98). Nonetheless, we encourage researchers to consider the question of sample size for such complex cases in the future.

Discussion and Recommendations

In this article, we demonstrated the difficulties in calculating sample size for a matched experimental audit. Researchers typically approach power and sample size from the perspective of magnitude differences. Yet, we show that for the matched audit design, phrasing the question in this manner is dependent on two factors. First, the total amount in the concordant cells contributes to power and sample size calculations. Second, the relative amount in the discordant cells requires a higher sample size as the difference between the cells decreases and the sum of the discordant cells approaches 1. These two factors

result in a scenario in which, for a given power, different sample sizes are required for the *same* magnitude difference between the treatment and the control. We also show that these difficulties are greatly complicated by an increase in either the number of treatments or the number of outcome categories. Thus, we encourage researchers and reviewers to approach power and sample size for audit designs with a *range* of potential magnitude differences in mind.

Given the importance of audit studies to the social sciences, as discussed in our review, we do not intend to diminish their utility. Instead, carefully designed audit studies have the potential to identify causal mechanisms in ways generally not possible through survey research. Rather, our goal has been to gather the various statistical tests for audit designs and to explicate the challenges of sample size selection to facilitate the design and effectiveness of these approaches. We provide functions in R to assist researchers planning an audit study. In consideration of both the importance of these studies and our analytic results, we offer the following recommendations regarding sample size calculations for those considering a matched audit design.

Recommendation 1: Conduct a Pilot Study for Sample Size Calculation Purposes

Although several audit articles mention a pilot or testing period, these often are geared toward fine tuning the audit procedure itself (e.g., appropriate signaling of the treatment on resumes and training of live testers). Once such issues are resolved, we highly recommend an additional pilot for purposes of computing sample size. One of the advantages of each of the sample size formulas presented is that they only require the *proportion* in each discordant cell. When the pilot experimental units are randomly selected from the same population as the planned actual experiment, the proportions in each pilot cell should approximate those expected for the larger sample. Even a pilot with a small sample size can provide the necessary proportions, which will be far more helpful than uninformed sample size calculations or extrapolation based on the past studies (as discussed subsequently).

An accurate sample size calculation based on a pilot period is also extremely useful for both grantors and grantees in the funding application process. Granting agencies typically expect sample size calculations for primary data collection. A more informed sample size calculation, such as that based on a pilot study, would give granting agencies greater confidence that a proposed budget is appropriately allocated. Grantees who have conducted a

pilot thus have the advantage of confidently demonstrating the minimum sample size (and corresponding budget) necessary to detect an experimental effect with an acceptable level of power.

Recommendation 2: Prepare for Several Different Concordant Sums

Of course, even pilot studies require some level of funding, potentially limiting this possibility. Another alternative is to use past studies, where available, to estimate the proportions beforehand. Whether using a pilot study or past studies, researchers should prepare for several different concordant sum possibilities, using the highest sample size. The concordant cells represent cases in which the testers receive the same outcome at the same experimental unit (e.g., both or neither job applicant receives a callback). The sum of these cells might be highly dependent on both location and time. For example, in an audit of hiring decisions, the ability of both or neither tester to receive an affirmative response from employers depends on prevailing economic conditions. Thus, past studies conducted in a different geographic location can only provide a rough estimate of the total concordant proportion, as both the strength of the local economy and the types of jobs available will differ across localities. Even estimates drawn from the same geographic location are likely to change, as the onset of recessions or shifts in sector-specific labor demand could alter the total cases where both or neither tester is called back. Moreover, such changes can also occur between the pilot study and the eventual granting of funds. For all of these reasons, researchers should consider a range of potential concordant sums, informed by contextual similarities and differences. For example, researchers might weight the results of past studies by industry to produce a distribution similar to the local economy that is being tested.

Recommendation 3: Prepare for Several Different Discordant Differences

Choosing the proportion in the discordant cells also presents challenges. Researchers must look beyond treatment/control differences, since sample size requirements are quite different for the same magnitude difference in the discordant cells (as shown on any left–right diagonal in Table 4). Our hypotheses typically assume that the experimental unit will prefer the control to the treatment when given an otherwise equal choice, such that it would be rare for the experimental unit to respond affirmatively to only the treatment. It may thus be hard to imagine why researchers must prepare for high values

for the case where the treatment receives an affirmative response, while the control does not (π_{01}). Yet, even in our empirical example, we used theory to prepare for this possibility. Theories of statistical discrimination posit that employers use race to draw "quick and dirty" assumptions about group differences in productivity and other characteristics, particularly when they lack detailed information about applicants (Arrow 1973; Bielby and Baron 1986; Braddock and McPartland 1987; Moss and Tilly 1996; Phelps 1972; Tomaskovic-Devey and Skaggs 1999). Thus, if employers assume applicants, particularly African Americans males, harbor a serious criminal record, disclosing an arrest might actually alleviate concerns based on statistical discrimination, causing the employer to favor the applicant with the arrest. We still viewed this as an overall less likely occurrence in our sample size calculations, but accounted for up to a 15 percent callback rate where the arrest treatment would receive a callback when the control did not.

In general, the weaker or "milder" the treatment, the more researchers might expect higher values in that cell. Although we had theoretical reasons to believe employers might only call our treatment tester, other similar studies might have no reason to assign high proportions to that cell in sample size calculations. For example, with Pager's (2003) treatment of a felony prison record, there is little reason to believe that divulging such a record would put an applicant in a favorable position over an equally qualified control applicant. In contrast, our own misdemeanor arrest treatment was considerably milder. Similarly, in Correll et al.'s (2007) test of parenthood, while we might expect nonparents only to receive a callback more of the time, we would not expect parenthood to be a wholly disqualifying characteristic. Thus, where the treatment is not wholly disqualifying, additional sources of variation that are not easily foreseeable or accountable might affect this cell. For example, an employer might possibly prefer parents (Correll et al. 2007), racial minorities (Pager, Western, and Bonikowski 2009), or those of a stigmatized religious identity (Wright et al. 2013), or the honesty associated with divulging a criminal record (Uggen et al. 2014).

Finally, we caution that additional sources of variation may occur in field experiments compared to laboratory experiments. Some of the cases in which only the treatment received a callback might be due to other nonrandom factors. For example, jobseekers who make direct contact are much more likely to be called back by employers, who may wish to provide a "second chance" to an otherwise promising applicant (Pager, Western, and Sugie 2009:206). If only the applicant with an arrest record is successful in making contact with the hiring authority, this may result

in a callback only to the treatment. In a bivariate test such as those presented here, this additional source of variation is not addressed. Although we emphasize the importance of collecting important covariates in audit studies, we note that a properly randomized study should remove covariation between the treatment and the covariates that are subject to the randomization process. Although the covariates can still have significant and sizable effects on the outcome, they should not affect the magnitude of the coefficient and standard error, and hence significance, of the treatment effect. For example, of the 18 covariates in our audit study, only 1 (contact) affected the treatment's coefficient and standard error to any discernable degree.[11] Thus, covariates are less a hindrance to power calculations in the experimental context, though they may be relevant to sample size if they affect the proportion in the discordant cells. Regardless of whether covariation exists between a covariate and the treatment, covariates can still covary with the outcome, and thus a multivariate framework still proves informative, as is presented in most sociological audit studies and discussed previously.

For many of the same reasons expressed concerning the sum of the concordant cells (e.g., labor market conditions), it is also difficult to determine how often the control will receive a callback when the treatment does not (π_{10}). For each of these reasons, as in our empirical example, we suggest the consideration of a range of possible discordant combinations.

Recommendation 4: Report Results at the Experimental Unit Level

In the absence of a pilot study, we recommended using past studies for the assumed proportions, a common practice in sample size and power analyses. Unfortunately, many empirical articles do not present results in a paired format such as those in Table 6 (see Riach and Rich 2002 for a summary of economics articles in this format), but rather only present the marginals. Such an omission does not imply any statistical limitations of past studies, but rather implies the absence of a uniform standard for presentation. This is understandable, given the novelty of audit studies within sociology, as well as possible trade-offs between presentation and replicability. For the sake of future designs, however, we highly encourage authors to present the results in the paired format, utilizing the tests presented previously. Future researchers can then apply those values as an estimate for their proportions, and consider an appropriate range around them. By referencing multiple studies on similar topics, those designing audits can best calibrate that range. For example, in terms of audit studies

testing criminal records, a future researcher might consider our empirical example on a misdemeanor arrest and Pager's (2003) study on a felony conviction with time served as the minimum and maximum proportions they should consider. Armed with this information, more informed sample size decisions can be made. If no research has been conducted for the specific treatment or experimental unit under consideration, a pilot study may be more imperative.

Recommendation 5: Consider a Nonmatched Design

Finally, where cost is a major consideration, we encourage authors to consider a nonmatched experimental design. This would involve only sending one randomly selected tester to each experimental unit. In his influential critique of field experiments, Heckman (1998) argues exactly such a case, stating that the use of matched pairs is not necessarily preferable to sending randomly assigned testers to *different* job sites. Even when only one tester is sent to each experimental unit, the randomization process, here achieved through random allocation of the units to treatment and control rather than matching and random selection in the case of an audit, should ensure no systematic bias in the assignment of the treatment (Cox 1958). Substantively, the matched approach is not necessary unless one wishes to account for *within*-employer effects (e.g., by including random experimental-unit intercepts to account for concordance; see Agresti 2002:410–11, 467–68, 493–501). Of course, researchers may very well have such an interest in within-employer effects.

Despite yielding results with true experimental rigor, the relevant question here is whether there is a sample size advantage to one approach or the other. As described previously, Donner and Li (1990) showed that the independent Pearson's χ^2 test is related to the matched tests presented via a weight that measures intraclass correlation. This simple connection shows that lower sample sizes are required for matched tests when there is a greater expected degree of concordance (as was expected and occurred in our audit). Conversely, lower sample sizes are required for independent tests where there is a greater expected degree of discordance.[12] This result makes perfect sense: When there is no effect of the experimental unit, it is irrelevant whether one sends testers to the same unit. The implication for sample size is straightforward. When researchers expect great concordance within experimental units, a matched design would economize funds. When researchers expect a high degree of discordance, an independent design would be more cost effective, though one should still consider substantive interests in the effect of the experimental unit.

Conclusion

Experimental audit studies represent a powerful and flexible tool for drawing causal inferences about social processes. In all such studies, the power to detect a certain magnitude difference between the groups represents a key design consideration. Unfortunately, the paired nature of the data complicates efforts to compute a priori power calculations. This article has presented sample size and power calculations from our own empirical study, then offered formulas and examples for cases involving more than two treatments (Cochran's Q test) and nominal outcomes (Stuart–Maxwell test). Beyond gathering these tests and explicating the challenges of sample size selection, we offer both concrete recommendations and the pertinent functions in R for estimating the appropriate size of future audit studies.

Authors' Note

The R functions referenced herein, including instructions for use, are publicly available on the first author's website at mikevuolo.com.

Acknowledgments

We are grateful to Zack Almquist and Sarah Mustillo for feedback on previous versions and to Lindsay Blahnik and Emily Harris for editorial assistance.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: 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). Uggen is additionally supported by a Robert Wood Johnson Health Investigator Award.

Notes

1. This table builds upon a similar presentation by Pager (2007), but expands the scope beyond racial discrimination and updates the list to include studies published in the ensuing seven years. Our search was conducted in August 2014. As with any indicator of scholarly influence, of course, Google Scholar can only provide an approximation of a publication's impact.

2. Importantly, audit tests are certainly not unique in conforming to a generalized Cochran–Mantel–Haenszel (CMH) test. As described previously, the 2×2 case is utilized in the medical literature in clinical trials. For example, the same patient may be administered both a drug and a placebo on separate occasions, and subsequently tested for a positive response (Royston 1993). The generalized CMH test can also be applied to survey research. For example, in describing McNemar's test, Agresti (2002:409–11) used the example of approval versus disapproval of a prime minister repeated among the same respondents on two occasions. We note, however, that for survey research, while providing a preliminary bivariate test, a CMH test would not provide the same level of scientific evidence of a relationship that results from an experiment. For this reason and those outlined in the introduction (increasing popularity, scientific yield, unique recommendations, cost and design considerations, and increasing complexity not present in other disciplines), we focus on audit studies, but note that the formulas and functions provided here can also be applied to these other disciplines and designs.
3. Additional stratification variables are discussed in the Cautions for Extensions of the Method section.
4. Additional explanatory variables of interest that do not vary within employer (e.g., race in Pager [2003]; gender in Correll et al. [2007]), as well as other covariates, are discussed in the Cautions for Extensions of the Method section.
5. We use the terminology "one-sided" and "two-sided" as referring to the hypothesis. We explicitly avoid the use of the word "tail" as even a two-sided χ^2 test refers to probabilities in the upper-tail only. For a χ^2 test, a one-sided test shifts the critical value downward and assigns more value to the upper tail, while a two-sided test has a higher critical value and assigns less value to the upper tail (rather than shifting some of the critical region to the lower tail). In the case of more than two treatments or outcomes, we present the two-sided hypothesis since the directionality is less straightforward.
6. Of course, researchers should be sensitive to asymptotic properties of the statistical tests (for a discussion, see Park [2002]), where small sample sizes should be approached with caution.
7. The four testers were grouped into pairs by race with one White pair and one African American pair, selected on the basis of shared physical and personal characteristics. Since only the same race was sent to a given employer, each race constitutes a separate experiment rather than the interactive treatment of race by misdemeanor record (see Cautions for Extensions of the Method section). For the illustrative purposes of this article, we discuss the pooled results with misdemeanor record as the treatment. For a discussion of the results by race, including a discussion of sample size and power, see Uggen et al. (2014).

8. For the 2 × 2 case, the numbers for the marginal proportions in Table 6 are those under consideration in the null hypothesis. Since the treatment and control share the employers where both receive a callback, the percentage point difference will be identical between the marginals and the discordant cells. Treating those marginals as independent, however, would produce erroneous sample size calculations (as well as statistical test results). As the formulas show, one must distinguish the cases in which both received callbacks. When there are more than two treatments, this similar percentage point difference will no longer apply, as multiple proportions account for each marginal (see section, "Cochran's *Q* Test: More Than Two Repeated Measures").
9. Note that post hoc power calculations make the assumption that the magnitude difference observed in the sample is the same as that of the population. See Hoenig and Heisey (2001) for cautions concerning post hoc power calculations using sample data. We only discuss these differences for illustrative purposes.
10. An alternative to the Stuart–Maxwell test is Bhapkar's test (1966). The difference between them is the formula used to estimate the covariance matrix. Asymptotically, the two tests are equal and Ireland, Ku, and Kullback (1969) provide a formula that directly relates the two tests. Bhapkar's test works better with small samples. Since they are asymptotically similar and the Stuart–Maxwell test is the direct generalization of McNemar's test, we present those formulas and calculations.
11. The covariates tested included contact, race, monthly unemployment, presence of minority employees, tester order, online versus paper ad source, test order over the course of the study, and industry, as well as the neighborhood-level characteristics of percentage under 18 years old, African American, same residence as prior year, single-headed households, who speak English less than very well, with bachelor's degree aged 25 and older, below poverty, adults employed, voting Democrat, and median household income and index I crime rate (models with covariates available upon request). Note that most of the covariates are characteristics of the employer, including its location, which is the object to which the randomization process is applied, which prevents covariation with the treatment effect. On the other hand, contact with the hiring authority, the one covariate that did covary with the treatment, is not easily subject to the randomization process (e.g., arriving at a restaurant during lunch might hinder an applicant's ability to make contact with the hiring authority). Note that contact is not an issue in correspondence studies.
12. As described in the section for Cochran's *Q*, concordance and discordance are less straightforward beyond the case of two treatments. Thus, researchers should not only consider concordance in terms of cases where all testers receive the same response but also the cases where the majority or higher of testers received

the same response. That is, there are different degrees of concordance that must be considered in this decision.

Supplemental Material

The online appendices are available at http://smr.sagepub.com/supplemental.

References

Agresti, Alan. 2002. *Categorical Data Analysis*. 2nd ed. Hoboken, NJ: Wiley.

Arrow, Kenneth J. 1973. "The Theory of Discrimination." Pp. 1-33 in *Discrimination in Labor Markets*, edited by O. C. Ashenfelter and A. Rees. Princeton, NJ: Princeton University Press.

Ayres, Ian. 1991. "Fair Driving: Race and Gender Discrimination in Retail Car Negotiations." *Harvard Law Review* 104:817-72.

Ayres, Ian and Peter Siegelman. 1995. "Gender and Race Discrimination in Bargaining for a New Car." *American Economic Review* 85:304-21.

Bendick, Marc, Jr., Lauren E. Brown, and Kennington Wall. 1999. "No Foot in the Door: An Experimental Study of Employment Discrimination against Older Workers." *Journal of Aging & Social Policy* 10:5-23.

Bendick, Marc, Jr., Charles W. Jackson, and Victor A. Reinoso. 1994. "Measuring Employment Discrimination through Controlled Experiments." *The Review of Black Political Economy* 23:25-48.

Bendick, Marc, Jr., Charles W. Jackson, Victor A. Reinoso, and Laura E. Hodges. 1991. "Discrimination against Latino Job Applicants: A Controlled Experiment." *Human Resource Management* 8:436-55.

Bendick, Marc, Jr., Charles W. Jackson, and J. Horacio Romero. 1997. "Employment Discrimination against Older Workers: An Experimental Study of Hiring Practices." *Journal of Aging & Social Policy* 8:25-46.

Berk, Richard A., Alec Campbell, Ruth Klap, and Bruce Western. 1992. "The Deterrent Effect of Arrests in Incidents of Domestic Violence: A Bayesian Analysis of Four Field Experiments." *American Sociological Review* 57:698-708.

Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94:991-1013.

Bhapkar, V. P. 1966. "A Note on the Equivalence of Two Test Criteria for Hypotheses in Categorical Data." *Journal of the American Statistical Association* 61: 228-35.

Bielby, William T. and James N. Baron. 1986. "Men and Women at Work: Sex Segregation and Statistical Discrimination." *American Journal of Sociology* 91: 759-99.

Birch, M. W. 1965. "The Detection of Partial Association, II: The General Case." *Journal of the Royal Statistical Society* B 27:111-24.

Braddock, Jomills Henry and James M. McPartland. 1987. "How Minorities Continue to Be Excluded from Equal Employment Opportunities: Research on Labor Market and Institutional Barriers." *Journal of Social Issues* 43:5-39.

Chow, Shein-Chung, Jun Shao, and Hansheng Wang. 2003. *Sample Size Calculations in Clinical Research*. New York: Marcel Dekker.

Cochran, W. G. 1950. "The Comparison of Percentages in Matched Samples." *Biometrika* 37:256-66.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20:37-46.

Correll, Shelley J., Stephen Benard, and In Paik. 2007. "Getting a Job: Is There a Motherhood Penalty?" *American Journal of Sociology* 112:1297-339.

Cox, David R. 1958. *Planning of Experiments*. Hoboken, NJ: Wiley.

Donner, Allan. 1992. "Sample Size Requirements for Stratified Cluster Randomizations Designs." *Statistics in Medicine* 11:743-50.

Donner, Allan and K. Y. Robert Li. 1990. "The Relationship between Chi-square Statistics from Matched and Unmatched Analyses." *Journal of Clinical Epidemiology* 43:827-31.

Drydakis, Nick. 2009. "Sexual Orientation Discrimination in the Labour Market." *Labour Economics* 16:364-72.

Esmail, A. and S. Everington. 1993. "Racial Discrimination against Doctors from Ethnic Minorities." *British Medical Journal* 306:691-92.

Esmail, A. and S. Everington. 1997. "Asian Doctors Are Still Being Discriminated against." *British Medical Journal* 306:1619.

Feins, Judith D. and Rachel G. Bratt. 1983. "Barred in Boston: Racial Discrimination in Housing." *Journal of the American Planning Association* 49:344-55.

Fleiss, Joseph L. 1981. *Statistical Methods for Rates and Proportions*. New York: Wiley.

Heckman, James J. 1998. "Detecting Discrimination." *Journal of Economic Perspectives* 12:101-16.

Hoenig, John M. and Dennis M. Heisey. 2001. "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis." *The American Statistician* 55: 19-24.

Ireland, C. T., H. H. Ku, and S. Kullback. 1969. "Symmetry and Marginal Homogeneity of an r $\times$ r Contingency Table." *Journal of the American Statistical Association* 64:1323-41.

Kenney, Genevieve M. and Douglas A. Wissoker. 1994. "An Analysis of the Correlates of Discrimination Facing Young Hispanic Job-seekers." *American Economic Review* 84:674-83.

Lachenbruch, Peter A. 1992. "On the Sample Size for Studies Based upon McNemar's Test." *Statistics in Medicine* 11:1521-25.

Lachin, John M. 1977. "Sample Size Determinations for r × c Comparative Trials." *Biometrics* 33:315-24.

Lageson, Sarah, Mike Vuolo, and Christopher Uggen. 2014. "Legal Ambiguity in Managerial Assessments of Criminal Records." *Law & Social Inquiry*. Online early.

Lahey, Joanna N. 2008. "Age, Women, and Hiring: An Experimental Study." *Journal of Human Resources* 43:30-56.

Landis, J. Richard, Eugene R. Heyman, and Gary G. Koch. 1978. "Average Partial Association in Three-way Contingency Tables: A Review and Discussion of Alternative Tests." *International Statistical Review* 46:237-54.

Mantel, Nathan and David P. Byar. 1978. "Marginal Homogeneity, Symmetry and Independence." *Communications in Statistics—Theory and Methods* 7:953-76.

McNemar, Quinn. 1947. "Note on the Sampling Error of the Difference between Correlated Proportions or Percentages." *Psychometrika* 12:153-57.

Moss, Philip and Chris Tilly. 1996. "'Soft' Skills and Race: An Investigation of Black Men's Employment Problems." *Work and Occupations* 23:252-76.

Neumark, David, Roy J. Bank, and Kyle D. Van Nort. 1996. "Sex Discrimination in Restaurant Hiring: An Audit Study." *Quarterly Journal of Economics* 111:915-41.

Newcombe, Robert G. 1996. "The Relationship between Chi-square Statistics from Matched and Unmatched Analyses." *Journal of Clinical Epidemiology* 49:1325.

Pager, Devah. 2003. "The Mark of a Criminal Record." *American Journal of Sociology* 108:937-75.

Pager, Devah. 2007. "The Use of Field Experiments for Studies of Employment Discrimination: Contributions, Critiques, and Directions for the Future." *The Annals of the American Academy of Political and Social Science* 609:104-33.

Pager, Devah and Lincoln Quillian. 2005. "Walking the Talk? What Employers Say Versus What They Do." *American Sociological Review* 70:355-80.

Pager, Devah, Bruce Western, and Bart Bonikowski. 2009. "Discrimination in a Low-wage Labor Market a Field Experiment." *American Sociological Review* 74:777-99.

Pager, Devah, Bruce Western, and Naomi Sugie. 2009. "Sequencing Disadvantage: Barriers to Employment Facing Young Black and White Men with Criminal Records." *The Annals of the American Academy of Political and Social Science* 623:195-213.

Park, Taesung. 2002. "Is the Exact Test Better than the Asymptotic Test for Testing Marginal Homogeneity in 2 × 2 Tables?" *Biometrical Journal* 44:571-83.

Patil, Kashinath D. 1975. "Cochran's Q Test: Exact Distribution." *Journal of the American Statistical Association* 70:186-89.

Phelps, Edmund. 1972. "The Statistical Theory of Racism and Sexism." *The American Economic Review* 62:659-61.

Pinheiro, Jose C. and Douglas M. Bates. 2004. *Mixed-effects Models in S and S-PLUS*. New York: Springer.

Quillian, Lincoln. 2006. "New Approaches to Understanding Racial Prejudice and Discrimination." *Annual Review of Sociology* 32:299-328.

R Development Core Team. 2006. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Retrieved (http://www.R-project.org).

Riach, P. A. and J. Rich. 1987. "Testing for Sexual Discrimination in the Labour Market." *Australian Economic Papers* 26:165-78.

Riach, P. A. and J. Rich. 1991. "Testing for Racial Discrimination in the Labour Market." *Cambridge Journal of Economics* 15:239-56.

Riach, P. A. and J. Rich. 2002. "Field Experiments of Discrimination in the Market Place." *The Economics Journal* 112:F480-518.

Rosner, Bernard. 2011. *Fundamentals of Biostatistics*. 7th ed. Boston, MA: Brooks/Cole.

Royston, Patrick. 1993. "Exact Conditional and Unconditional Sample Size for Pair-matched Studies with Binary Outcome: A Practical Guide." *Statistics in Medicine* 12:699-712.

Scott, William A. 1955. "Reliability in Content Analysis: The Case of Nominal Scale Coding." *Public Opinion Quarterly* 19:321-25.

Sherman, Lawrence L. and Richard A. Berk. 1984. "The Specific Deterrent Effects of Arrest for Domestic Assault." *American Sociological Review* 49:261-71.

Sherman, Lawrence L., Denise C. Gottfredson, Doris L. MacKenzie, John Eck, Peter Reuter, and Shawn D. Bushway. 1998. *Preventing Crime: What Works, What Doesn't, What's Promising*. National Institute of Justice Research in Brief. Washington, DC: USGPO.

Stuart, Alan. 1955. "A Test for Homogeneity of the Marginal Distributions in a Two-way Classification." *Biometrika* 42:412-16.

Tilcsik, András. 2011. "Pride and Prejudice: Employment Discrimination against Openly Gay Men in the United States." *American Journal of Sociology* 117: 586-626.

Tomaskovic-Devey, Donald and Sheryl Skaggs. 1999. "An Establishment Level Test of the Statistical Discrimination Hypothesis." *Work and Occupations* 26:420-43.

Turner, Margery Austin, Michael Fix, and Raymond J. Struyk. 1991. *Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring*. New York: Urban Institute.

Turner, Margery Austin and Maris Mikelsons. 1992. "Patterns of Racial Steering in Four Metropolitan Areas." *Journal of Housing Economics* 2:199-234.

Uggen, Christopher, Mike Vuolo, Sarah Lageson, Ebony Ruhland, and Hilary Whitham. 2014. "The Edge of Stigma: An Experimental Audit of the Effects of Low-level Criminal Records on Employment." *Criminology* 52:627-54.

Wallenstein, Sylvan and Agnes Berger. 1981. "On the Asymptotic Power of Tests Comparing K Correlated Proportions." *Journal of the American Statistical Association* 76:114-18.

Weichselbaumer, Doris. 2003. "Sexual Orientation Discrimination in Hiring." *Labour Economics* 10:629-42.

Wright, Bradley R. E., Michael Wallace, John Bailey, and Allen Hyde. 2013. "Religious Affiliation and Hiring Discrimination in New England: A Field Experiment." *Research in Social Stratification and Mobility* 34:111-26.

Yinger, John. 1986. "Measuring Discrimination with Fair Housing Audits: Caught in the Act." *American Economic Review* 76:881-93.

Yinger, John. 1991. "Acts of Discrimination: Evidence from the 1989 Housing Discrimination Study." *Journal of Housing Economics* 1:318-46.

Author Biographies

Mike Vuolo is an Assistant Professor of sociology at Purdue University. In addition to statistics and methodology, his substantive research interests include crime, law, and deviance, sociology of work and education, health, substance use, and the life course.

Christopher Uggen is Distinguished McKnight Professor of Sociology and Law at the University of Minnesota. He studies crime and social inequality, firm in the belief that good science can light the way to a more just and peaceful world.

Sarah Lageson is a PhD candidate in sociology at the University of Minnesota and will begin as an assistant professor of criminal justice at Rutgers University-Newark in 2015. She studies law, crime, technology, and media. Her research examines the growth and effects of criminal histories and crime data on the Internet.