Using Random Telephone Sampling to Recruit Generalizable Samples for Family Violence Studies

Amy M. Smith Slep, Richard E. Heyman, Mathew C. Williams, Cheryl E. Van Dyke, and Susan G. O’Leary
Stony Brook University

Convenience sampling methods predominate in recruiting for laboratory-based studies within clinical and family psychology. The authors used random digit dialing (RDD) to determine whether they could feasibly recruit generalizable samples for 2 studies (a parenting study and an intimate partner violence study). RDD screen response rate was 42–45%; demographics matched those in the 2000 U.S. Census, with small- to medium-sized differences on race, age, and income variables. RDD respondents who qualified for, but did not participate in, the laboratory study of parents showed small differences on income, couple conflicts, and corporal punishment. Time and cost are detailed, suggesting that RDD may be a feasible, effective method by which to recruit more generalizable samples for in-laboratory studies of family violence when those studies have sufficient resources.

**Keywords:** sampling, generalizability, random digit dialing, family violence, intimate partner violence

The viability of studying samples to learn about populations is entirely dependent on the generalizability (either demonstrated or inferred) of the samples studied. In many areas of psychology, however, the tradition of using convenience samples is longstanding, and the representativeness of samples often is not evaluated. This state of affairs may compromise the validity of results in psychological research and certainly complicates the interpretation of inconsistent findings when they arise.

Representative sampling for intensive (i.e., laboratory, not survey) research in clinical psychology rarely occurs. Investigators often recruit volunteers from among people seeking treatment for the problem of interest (e.g., Cascardi & Vivian, 1995), even if treatment is not a factor of interest in the study. When investigators are interested in generalizing to the population with regard to the problem of interest (e.g., all child-abusive parents), not just the population of those seeking or mandated to intervention services, this strategy is unlikely to produce results that are representative of the population. Alternatively, investigators advertise for participants and screen those who volunteer for the problem of interest (e.g., Margolin, John, & Gleberman, 1988). Although this strategy often generates the needed volunteers, it presents several problems: (a) It is difficult to know how well the sampled population (e.g., those reading a free weekly newspaper) represents the full population; (b) the convenience sample may differentially represent volunteers from problem and nonproblem populations, as a result of differences in perceived stigma or the desire to avoid scrutiny; (c) convenience sampling will especially hinder investigators seeking answers to clinical questions that rely on representative proportionality (e.g., What proportion of children who witness partner violence have behavioral problems? How often is women’s physical partner aggression in self-defense? What is the relative importance of different risk and protective factors?).

Fortunately, several methods for identifying and recruiting what are hopefully more generalizable samples are now being applied in psychology studies (e.g., Baucom, Epstein, Rankin, & Burnett, 1996; Hughes, Gordon, & Gaertner, 2004; Kroff, 1987). Among these approaches, random digit dialing (RDD) methods\(^1\) are particularly promising. RDD methods are most often used in large-scale survey

---

\(^1\) We say “methods” because, as Taylor (1997) noted, there is variability in the ways in which even major survey forms carry out random digit dialing.
research but are also common in epidemiological and medical studies (e.g., Gilliland, Li, Baumgartner, Crumley, & Samet, 2001; Rothman, Stanford, Kuniyuki, & Berger, 2004). Telephone saturation is near complete within American families (Casady & Lepkowski, 1998), and telephone interviews can quickly and efficiently identify individuals or families fitting particular criteria (e.g., Bunin, Baumgarten, Norman, Strom, & Berlin, 2005; Hartge et al., 1984). RDD remains a prime method of choice in epidemiological survey and case-control studies (Casady & Lepkowski, 1998), despite a variety of alternative cost-effective methods (e.g., mail and Internet surveys, Dilman, 2000; use of screened commercial telephone databases, Olson, Mignone, & Harlap, 2000; see also Ross, Spector, Olshan, & Bunin, 2004, for a discussion of the strengths and weaknesses of several other methods) and pitfalls (e.g., screening of calls with caller ID or answering machines, mobile phones replacing landlines).

It is well established that RDD produces very generalizable samples for telephone surveys (e.g., Hartge et al., 1984), but it is less clear how well RDD works in recruiting generalizable samples for studies in which data are collected in laboratories and how it can feasibly be implemented. Kroff’s (1987) used RDD as a recruitment strategy to obtain more representative samples for laboratory family research and compared the demographics of his study participants with those from published couple observation studies. Kroff used a multistage RDD recruitment strategy. Stage 1 consisted of an RDD survey including measures of relationship functioning; participants were not told that the survey was part of a larger endeavor. Stages 2 and 3 involved recruiting participants for a laboratory-based procedure. Participants recruited for the laboratory study by the multistage RDD method nearly matched the U.S. population on a variety of demographics, whereas the haphazardly recruited volunteers in the published literature were significantly younger, married for less time, and better educated, and they had fewer children.

Although Kroff’s (1987) findings bode well for RDD as a potential recruitment tool, further investigation of the usefulness of RDD-based laboratory study recruitment is needed. First, Kroff’s (1987) procedure for identifying and recruiting participants was multi-staged and fairly intensive. The RDD interview was a self-contained survey that made no mention of the observational study or the incentives available for the latter (e.g., participant payment). Survey respondents who qualified for the in-laboratory study were identified and mailed fliers about the study. Those who expressed interest then participated in an in-home informational session. Overall, this strategy worked fairly well: Although people who expressed interest in response to the mailed flyer, in comparison with those who did not express interest, differed in age and length of marriage, they did not differ with respect to marital adjustment. However, this intensive process exceeds the resources typically available. Thus, testing RDD as a direct participant identification/recruitment tool (i.e., without the letters and home meetings and with explicit mention of the larger study and its incentives) for family research is still needed. Second, Kroff’s (1987) phone survey did not contain questions about sensitive topics such as aggression; thus, neither the representativeness of the in-laboratory sample on such topics nor RDD’s utility for identifying such groups could be tested. Third, Kroff’s data were collected nearly 20 years ago in a small, mainly rural community in Illinois; contemporary work conducted in a more densely populated metropolitan area would provide useful additional information about the current viability and generalizability of this strategy. Finally, because response rates for RDD surveys generally have been falling (e.g., Curtain, Presser, & Singer, 2005), assessing the current viability of RDD as a recruiting strategy is in order.

Thus, many questions remain: Are randomly telephoned adults who are willing to answer questions specifically to identify eligible participants for an in-laboratory study different from the general population? Are those who are willing to come to a university for several hours, be videotaped, and complete extensive and intrusive assessments of family functioning different from those who are willing to participate in a telephone survey but not a laboratory-based study?

If self-selection biases are substantial, the resulting laboratory sample may be so dissimilar from the generalizable RDD sample that benefits of initially recruiting participants through RDD could be lost. It is also reasonable to speculate that RDD methods would be effective in recruiting generalizable samples of moderate- to high-functioning families (e.g., happy families with no history of aggression) but do not result in generalizable samples of problem families (e.g., distressed families with ongoing severe aggression). Finally, even if data suggest that samples recruited through RDD methods are generalizable, at what cost is the generalizability obtained? If recruiting samples through RDD is nearly as effortful as collecting the data of interest (which could be argued in the cases of Kroff, 1987) or must be contracted out to survey research firms (as most RDD efforts, including Kroff, 1987, are), then it may only be feasible to use RDD methods when sufficient resources are available and generalizability is of paramount importance.

Thus, we were interested in determining (a) the representativeness of study participants recruited with RDD methods and (b) the feasibility of using RDD to recruit community and clinical families to participate in laboratory-based studies. To do this, we evaluated data from a single RDD survey that was used to identify potential participants for two studies of family violence. The first focused on combined risk models of parent-to-child and partner aggression (the “family study”). Because this study included epidemiological questions and involved structural equation modeling as an analytic technique, it required a large (desired $N = 450$ laboratory couples) representative sample of families. Participants had to be living as a couple for at least a year, parenting a 3- to 7-year-old child who was the biological offspring of at least one of the parents, and able to complete questionnaires in English. The second study focused on anger and conflict processes in couples who were living
together for at least 1 year or who were married. It used a 2 × 2 design, crossing male-to-female aggression with relationship satisfaction (the “couple study”). To ensure distinct cells, couples in the nonaggressive cells had to report no history of male-to-female physical aggression in the history of the relationship; those in the aggressive cells had to report two or more acts of male-to-female aggression in the previous 12 months. Distressed couples needed to fall in the clinical range of relationship satisfaction, and happy couples needed to fall above the normative mean. Sample sizes of 50 were desired in each cell (desired N = 200 laboratory couples). Both studies required couples to spend 6 hr in the lab and be paid $250.

A multi-stage approach was used to evaluate the representativeness of the final samples for each study. First, demographics of the entire RDD sample (i.e., all phone respondents regardless of in-laboratory study eligibility) were compared with 2000 U.S. Census data for the county sampled in this study. Second, participants in each study were compared on demographics and family functioning measures with individuals who qualified for that study but did not participate. We did not hypothesize any differences because we expected our phone survey sample and in-laboratory study samples to be fairly representative. To facilitate evaluation of the feasibility of RDD methods, we describe our methods in detail. We also provide information about personnel time, calls made, and the approximate cost of our RDD efforts.

Method

Participants

Table 1 presents the final telephone and laboratory dispositions of all calls made with respect to the family and couple studies (see Ezzati-Rice et al., 2000, for a detailed description of the disposition categories and the derivation of the formulas for reporting response rates in a screening survey). Note that the response rates (45% for the family study and 42% for the couple study) applies to the simple response rate formula that takes into account an estimation of the eligibility rate of those who cannot be contacted), which produces rates 12–19% lower than the simple response rates in a screening survey. Note that the response rates (45% for the family study and 42% for the couple study) applies to the simple response rate formula that takes into account an estimation of the eligibility rate of those who cannot be contacted, which produces rates 12–19% lower than the simple response rates no longer considered appropriate.

Measures

Phone screen. The RDD phone screen was conducted by undergraduates at Stony Brook University. The script began, “Hello, my name is __________. I’m calling from Stony Brook University. We’re not trying to sell you anything. We’re randomly calling homes on Long Island to find people who might qualify to participate in our study of how families cope with conflict. Families who qualify for the study will be paid $250 for coming to the university for two 3-hour interviews. I’d like to take you 1 to 2 minutes of your time to ask you some questions to see if you qualify for the study. This survey is totally anonymous. And, of course, you don’t have to answer any question that you don’t want to.”

The first section (up to 15 questions) asked if there was at least one child between 3 and 7 years of age living in the household (and if “yes,” the ages of all children living in the household), marital/cohabitation status, whether the respondent and/or his or her partner was the child’s natural parent, ability of partner to speak English, ethnic/racial classification and age of participant and partner, family income in categories, and zip code. This section included all demographic and initial screening questions.

The second section (up to 21 questions) included items about the couple relationship and (for parents of 3- to 7-year-old children) parenting. Specifically, questions included the Quality of Marriage Inventory (QMI; Norton, 1983), an item about the frequency of arguments, questions about parent–child difficulties, five corporal punishment questions (spank child’s bottom with a bare hand; spank child’s bottom with an object; slap child’s hand, arm, or leg; pinch; shake) from the Parent–Child Conflict Tactics Scale (CTS-PC; Straus, Hamby, Finkelhor, Moore, & Runyan, 1998) that asked about the respondents’ and their partners’ parenting of the target child, five mild partner aggression questions (throw an object that could hurt, twist arm or hair, push or shove, grab, slap) from the Revised Conflict Tactics Scale (CTS2; Straus, Hamby, Boney-McCoy, & Sugarman, 1996) that asked about perpetration and victimization, and several questions about perceived ability to handle relationship and parenting problems.

Eligibility criteria. As stated above, to be invited to participate in the family study, a respondent had to report that (a) he or she or the partner was the biological parent of a child between the ages of 3 and 7 living in the household; (b) he or she had been cohabiting or married for at least 1 year; and (c) both the respondent and the partner could understand and read English.

To be invited to participate in the couple study, a respondent had to (a) report that he or she was married or had been cohabiting for at least 1 year, (b) report that both the respondent and the partner could understand and read English, and (c) meet both the relationship satisfaction criterion (i.e., happy or clinically distressed) and the male-to-female aggression criteria.

The relationship satisfaction criterion for the couple study used cut-points from the Dyadic Adjustment Scale (DAS; Spanier, 1976) that were converted into QMI scores on the basis of published formulas (Heyman, Sayers, & Bellack, 1994). To be considered “happy” on the phone screen, a respondent had to score at or above the DAS mean plus one standard error (as based on Eddy, Heyman, & Weiss, 1991), that is, a DAS score of 114 and a QMI equivalent of 37. To be considered “clinically distressed,” a respondent had to score at or below the traditional DAS criterion for distress minus one standard error (cf. Jacobson & Truax, 1991), or 92.5 and a QMI equivalent of 27. By adding the standard error, the phone screen was meant to be conservative; the laboratory criteria used the means only (i.e., QMI cut-points of 34 and 29, respectively). To be considered in an aggressive relationship, the respondent had to report two or more acts of male-to-female aggression during the past year; to be considered in a nonaggressive relationship, the respondent had to report no male-to-female aggression in the history of the relationship.

Procedures

Conducting the interviews. Interviewers were undergraduates earning research credits who worked for 1.5-hr periods. Training

2 For researchers interested in the consistency of telephone versus laboratory happiness by aggression groupings for the couple study, Supplemental Table A is available on the Web at http://dx.doi.org/10.1037/0893-3200.20.4.680.supp or http://psychology.psy.sunysb.edu/ftrlab/papers.html.
Table 1

Counts of Final Dispositions for the Family and Couple Studies

<table>
<thead>
<tr>
<th>Label</th>
<th>Disposition category</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>Total selected sample</td>
<td>229,106</td>
</tr>
<tr>
<td>D + E</td>
<td>Nonworking, out-of-scope, nonresidential</td>
<td>157,669</td>
</tr>
<tr>
<td>G + I</td>
<td>Noncontact, answering machine (working number, household status unknown), terminations without answering any questions</td>
<td>32,932</td>
</tr>
<tr>
<td>S1</td>
<td>Known household, screening for eligibility incomplete</td>
<td>3,866</td>
</tr>
<tr>
<td>S2</td>
<td>Likely householda</td>
<td>26,751</td>
</tr>
<tr>
<td>J</td>
<td>Screened household, not eligible according to study criteria</td>
<td>5,918</td>
</tr>
<tr>
<td>K</td>
<td>Screened household, eligible</td>
<td>1,970</td>
</tr>
<tr>
<td>L</td>
<td>Completed interview</td>
<td>1,785</td>
</tr>
<tr>
<td>M + N</td>
<td>Eligible household, incomplete interview or refusal</td>
<td>185</td>
</tr>
</tbody>
</table>

Response rateb
Total laboratory participation/final laboratory samplec

45% 42%
459/453 305/233

Note. Labels correspond to notation of Ezzati-Rice et al. (2000). n = 12,009 unique interviews ([J + K] = 2,645 [n who screened into both studies]).

a Call answered but household status determination could not be conducted or was incomplete.

b The response rate was calculated as follows:

Response rate = \( \frac{D + E + S1 + J + K}{\text{TOTAL}} \times \frac{J + K}{J + K + S1} \times \frac{L}{K} \)

c For the family study, 6 couples failed to complete both data collection sessions, were later found to have either not met inclusion criteria or skipped key questionnaires in the study. For the couple study, 72 couples failed to complete the protocol or were found to not qualify for any of the four groups on the basis of their in-laboratory reports.

consisted of six 1.5-hr sessions. The first two sessions were an orientation to RDD and the methods used for data recording/collection. In the third and fourth sessions, students were introduced to the telephone script and the nuances of doing clinical surveying. During the fifth and sixth sessions, students practiced the script through role playing. Before students were allowed to make actual phone calls, they were required to perform a role play almost flawlessly.

Six interviewers could work simultaneously (supervised by bachelor, master’s, or doctoral-level study staff), and approximately fourteen 1.5-hr interviewer periods occurred per week. An average of 164 calling periods occurred each semester. Calls took place Mondays through Thursdays from 2 p.m. to 3:30 p.m. and from 6 p.m. to 9 p.m., and Saturdays from 12 noon to 3 p.m. (not all of these periods were used across all sessions of calling). An average of 28.52 calls were made by interviewers during calling shifts. Averaging across all weeks during the semesters when RDD was occurring, including training time, 1,672 calls were made each week. Calls were only made when classes were in session. RDD began in September 1998 and ended in May 2002. The entire project required 12,053 hr of undergraduate research assistant time (including training).

Random digit dialing procedures. All scripts and procedures were approved by the university’s Institutional Review Board. Our RDD procedures were adapted in large part from those used in the 1985 National Family Violence Survey (Louis Harris & Associates, 1986). These procedures call for matching a central office code (the first three digits in a seven-digit local phone number) with a randomly generated four-digit line number (the final four digits in a seven-digit local phone number). We downloaded a list of all central office codes from the North American Numbering Plan Administration Website (http://www.nanpa.com/reports/reports_cocodes_assign.html) and eliminated all central office codes that were either (a) assigned to mobile phone providers3 or (b) located in Suffolk County, New York, but greater than a 45-min drive from the university. We identified 216 central office codes to be included in calling using these criteria. We created a grid in which each line consisted of (a) a randomly sorted central office code and (b) five randomly generated four-digit line numbers. We generated about 230,000 such lines and printed the grids for use by the telephone interviewers.

Interviewers manually dialed 229,106 numbers from the randomly generated list. For each row, interviewers used as many of the five 4-digit numbers as needed to conduct an interview. When the first call on that row resulted in a completed interview, the caller moved to the next row for the next call. If the first call on that row resulted in (a) a nonworking number, business, or no answer after four calls, or (b) refusal/hang-up (i.e., verbally refused to answer any questions, refused to continue once interview began, or hung up at any point), the caller stayed on the same row but next called the same central office code matched with the line number appearing in the next column. If the number dialed was a busy signal or a ringing but unanswered phone, the caller temporarily skipped this row and noted the time. The interviewer made another attempt to the same number a minimum of 15 min later. If the interviewer could not reach anyone at that number after four tries, he or she moved to the line number in the next column. If necessary, the interviewer repeated these procedures until all five columns were used or until an interview was completed. If the interviewer did not complete an interview but used all five columns, he or she moved to the next row. The status of each call made was written in a calling log and was later entered into a computerized database.

3 Our study was conducted prior to the introduction of “phone number portability,” the policy that allows a user to disconnect his or her land-line and use that phone number for his or her mobile phone. Thus, replicating our procedures now would invariably include a small number of mobile phones.
Eligible respondents, as identified from the phone screen, were invited to participate in the family study. If a respondent qualified for both the family and couple studies, he or she was informed of the couple study on completion of the family study. If a respondent qualified for the couple study only, he or she was invited to participate if the applicable cell (i.e., happy/never aggressive, distressed/never aggressive, distressed/aggressive, happy/never aggressive) was still open. To control for expected age differences in these four groups, the distressed/aggressive group served as the age criterion group. Age ranges (in 5-year increments) of the most prevalent group (happy/never aggressive) were sometimes closed to enrollment to preserve roughly equivalent age distributions. Thus, because of their overabundance, many otherwise qualifying happy/never aggressive respondents were told that they did not qualify. If a respondent was told that he or she qualified and he or she was interested in possibly participating, a principal investigator or project leader called him or her within the following few days to describe the study in detail, answer any questions, and schedule the couple for their laboratory visit(s).

The distribution of respondents who met telephone criteria for the couple study (i.e., including those who were screened out on account of age-matching restrictions) was as follows: happy/never aggressive (n = 4,528 [84.43% of qualifiers]); distressed/never aggressive (n = 178 [3.32% of qualifiers]); distressed/aggressive (n = 548 [10.22% of qualifiers]); and happy/aggressive (n = 57 [2.03% of qualifiers]).

**Results**

The first step in evaluating the representativeness of our RDD-recruited study samples was to compare all respondents who participated in a phone interview with all residents of Suffolk County, New York, according to the 2000 U.S. Census. A few compromises were necessary because of the structure of the census data. Using county data includes a wider catchment area than we sought to sample with the phone survey, which sought to call homes within a 45-min drive of the university. This radius eliminated the most eastern, and least populated, portion of the county from the survey catchment area. However, because central office codes do not perfectly map onto geographic location, analyses of the survey sample’s zip codes revealed that although the vast majority of respondents lived within the intended survey catchment area, some respondents did not, and the entire county was at least minimally represented in the final survey sample. Therefore, RDD data were compared with census data for the entire county.

To conduct these comparisons, we first examined the phone interview and census variables to identify demographic variables that could be compared. The following were asked (or could be recoded from questions asked) in the survey and were available in the census: Latino/Hispanic ethnicity of adults, number of races of adults, race of adults self-identifying as of one race, sex of adults, adult men’s ages, adult women’s ages, family type, and family income. For each comparison, we used all available survey data (regardless of whether that respondent was missing data on another variable).

Chi-square tests for goodness of fit were used to compare the frequency distributions of the RDD sample with those of the census. However, with very large samples, the chi-square statistic is highly sensitive to subtle differences between the observed and expected frequencies (Ferguson, 1966). Given that the RDD survey sample included over 12,000 interviews, with information about more than 24,000 adults, statistically significant differences between the RDD sample and 2000 census do not necessarily reflect meaningful differences. Therefore, effect sizes were computed to qualify the results of the chi-square tests. The appropriate index of effect size for a 2 × 2 contingency table is the phi coefficient (Φ), and Cramer’s Φ is the appropriate index of effect size for a contingency table larger than 2 × 2 (Aron & Aron, 1999). Cohen (1988) suggested that effects of approximately .10 are small, those equal to .30 are medium, and those exceeding .50 are large.

Observed and expected proportions are reported in Supplemental Table B, which is available on the Web. Comparisons revealed no differences for sex, χ²(1) = 3.91, Φ = .02, near zero effect size. Small, but statistically significant, differences emerged for Latino/Hispanic status (survey undersampled, χ²(1) = 182.40, Φ = .10, small effect size) and being of two or more races (survey undersampled, χ²(1) = 138.23, Φ = .09, small effect size). Significant differences with medium effect sizes were found for race (survey oversampled African Americans and American Indians and undersampled “some other race,” χ²(5) = 1,627.08, Cramer’s Φ = .31, medium effect size), men’s and women’s ages (survey oversampled people between ages of 30 and 44, men’s χ²(18) = 649.03, Cramer’s Φ = .28, women’s χ²(18) = 1,047.84, Cramer’s Φ = .34, medium effect sizes), family type (survey oversampled married couple households and undersampled single male and female households, χ²(2) = 1,180.07, Cramer’s Φ = .38, medium effect size), and family income (survey oversampled families with incomes between $30,000 and $100,000, χ²(5) = 342.09, Cramer’s Φ = .22, medium effect size).

The next step in evaluating the generalizability of the two study samples was comparing the final participants with people who met eligibility requirements for the two studies but ultimately did not participate. For the family study, Table 2 displays comparisons between laboratory participants (N = 453 couples) and RDD participants who were qualified to participate in the laboratory study but who did not (N = 1,362 RDD respondents) on all demographic and family functioning variables that were assessed in the phone interview. Twenty variables were available for comparison and are listed in Table 2. Prior to analyses, the distributions of continuous variables were examined for normality. Variables that showed significant deviations from normality (Fidell & Tabachnick, 2003) were transformed. When no transformations were necessary and when transformations sufficiently normalized the variables, continuous variables were compared with t tests. When transformations were not

---

4 This material is available in Supplemental Table B on the Web at http://dx.doi.org/10.1037/0893-3200.20.4.680.supp or http://psychology.psy.sunysb.edu/ftrlab/papers.html.
successful, nonparametric chi-square tests were used. Of the 20 variables compared, only four differed significantly between the two groups. The participants were less likely to report family incomes greater than $100,000, disagreed with their partners more frequently, and reported that both parents used a wider variety of corporal punishment. Although these differences were statistically significant, the effect sizes were all small.

We conducted similar comparisons for the couple study, that is, the final laboratory sample participants versus qualifying RDD respondents who did not participate in the laboratory study ($N = 4,526$). Because our factorial design set the targeted representation of couples in each of the four cells to be approximately even, which is markedly different from the groups’ natural representation in the population, these comparisons were conducted by group. Across the 13 variables tested within each of the four groups, laboratory participants differed from their corresponding group in the RDD pool only on the gender of the phone respondent $\chi^2(1) = 4.53, p < .05$. Distressed/aggressive respondent couples were significantly more likely to participate if the man, instead of the woman, was the phone respondent. The effect size was small.

Discussion

The purposes of this study were to address whether participants in laboratory studies recruited using random telephone sampling were reasonably representative of the larger population and to speak to the feasibility of this recruitment method in family research.

---

**Table 2**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Statistical comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex of respondent</td>
<td>$\chi^2(1) = 3.41$</td>
</tr>
<tr>
<td>Biological parent of target child</td>
<td>$\chi^2(1) = 0.93$</td>
</tr>
<tr>
<td>Marital status</td>
<td>$\chi^2(4) = 3.35$</td>
</tr>
<tr>
<td>Male respondents’ race</td>
<td>$\chi^2(4) = 1.40$</td>
</tr>
<tr>
<td>Female respondents’ race</td>
<td>$\chi^2(4) = 1.40$</td>
</tr>
<tr>
<td>Respondent age</td>
<td>$t(1733) = -1.30$</td>
</tr>
<tr>
<td>Partner age</td>
<td>$t(1733) = -1.62$</td>
</tr>
<tr>
<td>Family income</td>
<td>$\chi^2(5) = 13.29^*$</td>
</tr>
<tr>
<td>Number of adults in household$^a$</td>
<td>$\chi^2(8) = 7.77$</td>
</tr>
<tr>
<td>Relationship satisfaction$^b$</td>
<td>$t(1807) = -0.58$</td>
</tr>
</tbody>
</table>

How often do you and your partner have arguments?

How often do you have some kind of difficulty with [target child]?

What percentage of disagreements with your partner are about [target child]?

How difficult a child is [target child] to raise?

Variety of corporal punishments by respondent$^c$

Variety of corporal punishments by partner$^c$

Variety of acts of male-to-female physical aggression$^c$

Variety of acts of female-to-male physical aggression$^c$

How often do you and your partner solve your disagreements?

How successful are you in handling problems with [target child]?

How successful is your partner in handling problems with [target child]?

---

Note. Ns for the laboratory sample and qualifying nonparticipants, respectively, were 453 and 1,362.

$^a$ Untransformably skewed. $^b$ Reflected and inverse transformed. $^c$ Log transformed.

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

**Generalizability**

Are RDD respondents adequately representative of the population? Although our RDD survey, unlike most surveys, did nothing to select the sex of the RDD respondent, their sex was representative of distributions in the county’s population. For other demographic variables, not surprisingly, modest differences between RDD respondents and the population were observed. We slightly underrepresented respondents identifying themselves as Latino/Hispanic (perhaps because our study, and thus our RDD survey, was conducted only in English), those of some “other race,” and those of two or more races. We slightly overrepresented African Americans and American Indians, which can be seen as an advantage because our county underrepresents

---

Footnotes:

5 Only telephone screen status is known for qualifiers who did not participate in the laboratory study; thus, this comparison between those in the final laboratory sample and the telephone-only sample had to use telephone screen status for both groups. As shown in Supplemental Table A (available on the Web at http://dx.doi.org/10.1037/0893-3200.20.4.680.supp or http://psychology.psy.sunysb.edu/ftrlab-papers.html), the ns for telephone screen statuses for the final laboratory sample were as follows: happy/non-aggressive (97), distressed/non-aggressive (61), distressed/aggressive (33), and happy/aggressive (7). The 49 couples in the expanded criteria group did not cleanly fall in a telephone screening group and were therefore excluded from these analyses.

6 Space constraints precluded the publication of the full comparison table. It is available on the Web in Supplemental Table C at http://dx.doi.org/10.1037/0893-3200.20.4.680.supp or http://psychology.psy.sunysb.edu/ftrlab/papers.html.
these groups compared with national proportions. We slightly oversampled both married couple households and men and women 30 to 44 years of age (and undersampled unmarried singles) probably because the preamble to our survey described it as a way to qualify for a study of “how families cope with conflict” and perhaps because married people and 30- to 44-year-olds are more readily accessible by telephone. This was not problematic given the laboratory studies we were recruiting for, as singles were eliminated after the initial demographic screening questions anyway. Finally, we slightly oversampled middle-income families, probably because wealthy families are less swayed by possible monetary incentives and because poor and struggling persons may be less available by telephone (e.g., working several jobs). Overall our RDD survey sample was fairly representative of the county population, with many of the perturbations likely not affecting the representativeness of our targeted samples of couples parenting a child between 3 and 7 years of age and of couples married or living together.

Are RDD-recruited laboratory participants adequately representative of their respective populations? Because our RDD sample was fairly representative of the county population, it was appropriate to test whether our laboratory participants adequately represented all RDD-qualifying respondents on a number of variables for which there are no census data.

Individuals eventually participating in the family study, compared with all qualifying, nonparticipant parents in the RDD-screened pool, did not differ significantly on a host of substantive and demographic factors. Laboratory participants slightly deviated from RDD-only qualifiers in that they reported using a wider variety of corporal punishment and more frequent relationship conflict; they also were less likely to earn over $100,000 a year. The slight differences on family functioning variables might reflect that families with little conflict are less interested in contributing their time to a study of family conflict. The income difference may be because the monetary incentive was less motivating to this high-income group.

Couples eventually participating in the couple study, compared with their comparable RDD-qualifying nonparticipant counterparts (i.e., those in the same happiness by aggression cell), did not differ significantly on substantive or, with one exception, demographic variables. Distressed/aggressive couples were slightly more likely to have the man be the RDD-screened respondent; the effect size was small.

In conclusion, our laboratory samples were generally representative of the RDD sample, which in turn was fairly representative of the overall population. Thus, one can infer that RDD indeed results in laboratory samples that reasonably represent the populations from which they are drawn. Because Krokoft’s (1987) study clearly showed that haphazardly collected volunteer studies involving couples are highly discrepant from census data, whereas his RDD-generated sample was not, this study adds further evidence to the proposition that studies wishing to generalize to the larger general population (i.e., the desire of almost all non-treatment-oriented studies) should use RDD rather than convenience strategies to generate their pools of potential participants for intensive study.

Feasibility

We interviewed 12,009 households out of 229,106 randomly constructed numbers between September 1998 and May 2002. Note that we were using a fully manual system; we recently purchased an affordable computerized system (i.e., WinQuery 4.0; Morton, 2005) that automatically removes duplicate numbers from an investigator-loaded random phone number list; facilitates approximately seven times more calls per interviewer; and manages dialing, data entry, and scoring automatically. Despite the fact that fewer than approximately 5% of numbers resulted in an eligible household, RDD allowed us to identify (a) an abundance of qualifying, interested parents for the family study and (b) a plentiful group of happy/never aggressive and distressed/aggressive couples. As with other studies (e.g., Holtzworth-Munroen et al., 1992), recruiting distressed/never-aggressive couples was more challenging. Happy/aggressive couples were the most challenging; with a more efficient computerized calling strategy, it may be possible to locate more of these couples than was possible given our manual efforts. Overall, our success was likely improved by the large and sustained RDD efforts, involving recruitment for two studies and conducted in a densely populated area.

There were advantages of basing the RDD within our laboratory (as opposed to hiring an outside survey firm).
First, we spread the recruitment over several years, meeting the needs for a steady flow of couples through our laboratory. It is almost inconceivable, given the way survey firms operate, to imagine that an organization would stretch a project from a several-day or -week time frame to a several-year time frame without added expense. Alternatively, it would be useless to identify a large pool of potential participants and then require them to wait several years for their laboratory visit. Second, hiring a survey firm for the recruitment activities would have cost nearly as much as conducting both studies in their entirety. Relatedly, the staff hired to supervise RDD interviewers also served other roles in the study (e.g., conducting laboratory sessions, managing and analyzing data). Finally, we were able to open and close couple study groups on the basis of dynamic age matching and to make other adjustments tailored to the needs of the laboratory studies.

Finally, conducting RDD in our laboratory kept costs manageable (given that we had two grant-supported projects). We estimate the total cost of our RDD effort to have been approximately $87,000, considering that it occurred from September 1998 through May 2002. This is based on the actual salaries and fringe benefits for principal investigators, staff, and graduate students for the amount of effort they devoted to RDD, along with actual costs of phone calls and supplies. This cost translates to approximately $125 spent to recruit each in-laboratory study participant included in the final samples. The cost of identify-
ing individuals in less frequent subgroups, such as a couple with husband-to-wife aggression, is higher. We estimate that it cost approximately $450 to identify each couple who reported at least two acts of husband-to-wife aggression in the past year during the phone interview. Those costs are less if all couples eventually identified as meeting the aggression criteria on the basis of the laboratory assessment of both partners are included. Of course, involving unpaid undergraduate research assistants in the efforts likely helped keep costs down, although training and supervision costs were higher and productivity and response rates were lower (O’Rourke & Johnson, 1999) than they would have been if we had used professional interviewers. Many other factors, including the size, and therefore economy, of our efforts: our population-dense geographic location; and our decision to screen only one partner might have contributed to the relatively manageable cost and ease with which we recruited our samples. Although nearly $100,000 is not an insignificant amount of money, it does suggest that RDD recruitment should not be ruled out as too expensive to be feasibly implemented for laboratory study recruitment.

Limitations and Recommendations

Although we believe that the results of this study are quite encouraging regarding both the feasibility of RDD as a recruitment method and the generalizability of resulting samples, this study and the use of RDD as a recruitment tool are not without limitations. First, despite the relative crudeness of our procedures compared with the most expensive, extensive survey procedures (e.g., we did not mail out letters to potential recruits, pay for participation, employ “refusal converters,” use computer-assisted technologies, buy prescreened residential phone lists, or research non-contact numbers to ensure they were residential), our response rates (42–45% according to the conservative Massey/Ezatti-Rice formula) are comparable to those found in many of the more sophisticated studies (e.g., Curtain et al., 2005). Nevertheless, such response rates—even when coupled with fairly representative RDD—Census demographic comparisons—imply a nonresponse bias that may affect representativeness. Because of the difficulty of recruiting a perfectly representative sample by phone, some purposes might require using some of the methods described above and bearing the extra cost they entail.

Second, RDD was far less costly to conduct in our laboratory than it would have been to hire an outside contractor, but it was nonetheless a substantial undertaking. Researchers might consider using a single RDD survey to recruit for multiple studies (as we did) to minimize costs. Further, using a computerized system for dialing, determining qualification status, and tracking data substantially increases efficiency and reduces costs. It may also be cost-effective to purchase high-quality survey lists from firms that specialize in developing and prescreening these, depending on how quickly the user intends to complete their calling efforts. However, RDD may be beyond the reach of

most completely unfunded studies, particularly in research seeking low base-rate groups.

Second, introducing the phone survey as a way to identify potential participants for studies of how families cope with conflict appears to have affected the generalizability of the RDD survey sample. If laboratories were interested in collecting RDD survey data to answer general substantive questions (i.e., the survey was a stand-alone research project in itself), as well as identifying a select group to invite into a laboratory-based study, they should emphasize the survey research question as the purpose rather than focusing primarily on the subsample they might want to recruit.

Third, these findings indicate that RDD recruitment is not a panacea. Although findings suggest that both in-laboratory studies’ samples were reasonably representative, it is also the case that they were not perfectly generalizable. Thus, RDD can facilitate good generalizability, even among low—base-rate subpopulations, but it does not guarantee it. Finally, our results were no doubt partially a function of the population density of the area studied. Suffolk County, New York, includes nearly 1.5 million residents (2000 U.S. Census, 2003). The feasibility of using RDD to successfully recruit reasonably generalizable samples, especially of low—base—rate groups (e.g., happy aggressive couples), would likely be much different in a less populated area.

Whether Family Psychology in the 21st Century?

As researchers continue to struggle with balancing practicality and optimal methods, we strongly recommend explicit attention to issues of generalizability. Although pilot and other unfunded studies will continue to rely on cheap convenience samples, we demonstrated that funded studies can recruit fairly generalizable samples in a cost-effective manner. We are by no means the first family psychology study to attempt this. For example, studies of newlyweds (e.g., Leonard & Roberts, 1998) have sometimes used marriage license registries, and studies of children have long used birth records, as sampling frames. By thinking creatively about how to (a) improve the generalizability of our samples and (b) collect data that will allow us to evaluate and specify the generalizability of the final sample collected, family researchers will be able to move toward more solid scientific ground, even if convenience samples are still sometimes necessary.

For some reason, psychology seems to trail epidemiology in attending to the representativeness of participants in expansive protocols. Sometimes these participants are intended to represent a subset of the general population (in our case, parents of children between 3 and 7 years of age), and sometimes they are intended as controls in case-control medical research. Nevertheless, psychologists designing

---

7 We should note that although it is widely accepted that lower response rates indicate nonresponse bias, several studies have not found support for biasing effects (e.g., Curtain, Presser, & Singer, 2000; Keeter, Miller, Kohut, Groves, & Presser, 2000; Merkle & Edelman, 2002).
funded studies would do well to peruse the epidemiological literature (often available on the Medline, but not the PsyclINFO, abstract database) for guidance on a plethora of strategies that could yield more representative participants than the strategy of convenience sampling. For example, Ross, Spector, Olshan, and Bunin (2004) sketched the pros and cons of RDD; hospital, school, and cousin controls; and public records (e.g., births, marriages).

In our particular case, we believe that one of the major advantages of RDD was not that it guaranteed a representative sample but rather that it allowed us to collect data from individuals who completed the RDD screen but did not participate in the laboratory study. That important feature enabled us to examine the generalizability of the final in-laboratory sample on key variables of interest (e.g., relationship satisfaction, partner and parent–child physical aggression). If this were to become more typical in psychological research, it would likely be far easier to interpret apparently inconsistent findings in literatures. Our experience with RDD implies that when resources are available, “convenience” samples represent only a fleeting convenience; more robust, viable recruitment designs can provide not only a pool of interested laboratory participants but also important information on the generalizability of findings to the population of interest.

References


Received July 20, 2005
Revision received February 1, 2006
Accepted February 1, 2006