Address-Based Sampling for Socio-Demographic Studies of the U.S. Jewish Community

Zachary H. Seeskin[1], David Dutwin[1], Leonard Saxe[2]
[1]NORC at the University of Chicago
[2]Cohen Center for Modern Jewish Studies, Brandeis University

Abstract*
Jewish community studies, which provide estimates of the size and characteristics of Jewish populations in different metropolitan areas, provide critical information to inform planning and programming. Because of the small proportion of the population who identify as Jewish and the lack of official statistics on religion, sampling procedures are complex. Until recently, random digit dialing was considered the gold standard. Only recently has address-based sampling (ABS) been applied to Jewish community studies. Two recent scale applications of ABS are the 2020 Metropolitan Chicago Jewish Population Study and the 2021 Study of Jewish Los Angeles. Both designs employed dual-frame sampling from combined, deduplicated Jewish organization membership lists and a version of the U.S. Postal Service Computerized Delivery Sequence File licensed from a vendor. We discuss strategies critical for the success of implementing ABS for Jewish community studies, including using predictive modeling and geographic stratification to reach Jewish households. The paper considers the advantages, as well as limitations, of using ABS for the study of rare populations.

Key Words: Rare populations; hard-to-reach populations; multiple sampling frames; predictive modeling

1. Introduction

Jewish community studies are surveys conducted in local areas to provide estimates of the size and characteristics of Jewish populations. The studies provide critical information to local Jewish organizations for planning and programming, having particular importance because the U.S. Census and federal government surveys do not collect data about religious practice and background (Aronson, Boxer, & Saxe, 2016; Saxe, Tighe & Boxer, 2014). National and local organizations often need data on small or otherwise hard-to-reach subgroups of the Jewish community to better serve their needs. Among the subgroups Jewish community studies often seek to reach include families, young adults, denominational groups, interfaith households, the financially struggling, Jews of Color, Jews of different ethnic backgrounds, and LGBTQ Jews. Further, these studies often seek to provide geographic detail on the Jewish population for local organizations. An additional goal and challenge for these studies is to reach less engaged members of the Jewish community.

* Please direct correspondence about to Zachary Seeskin, Seeskin-Zachary@norc.org.

The fundamental challenge of conducting Jewish community studies is to survey a rare population and provide high-quality estimates within a reasonable budget, as Jews are estimated to be 2.4% of the U.S. population (Pew Research Center 2021). While new to the practice of Jewish community studies, address-based sampling (ABS) offers great promise for sample designs for Jewish community studies (Link et al. 2008, Harter et al. 2016). Historically, most Jewish community studies were conducted by list-assisted random digit dialing (RDD). In the current environment, however, RDD has notable weaknesses for Jewish community studies including increasing survey costs and lower population coverage of the households in a specific metropolitan area (Lavrakas et al. 2017). Among available sample design methodologies, ABS is particularly well-suited to conduct high quality Jewish community studies, providing the advantages of known probabilities of sample selection for every household in the catchment area and high population coverage at reasonable cost.

We discuss strategies for implementation of high-quality ABS design for Jewish community studies highlighting the approaches for two prominent recent studies, the 2020 Metropolitan Chicago Jewish Population Study (Aronson et al. 2021) and the 2021 Study of Jewish Los Angeles (Aronson et al. 2022). In discussing these methodologies, we highlight three practices for conducting ABS for Jewish community studies that supported these studies' success: (1) careful frame development and sample design utilizing many membership and participant lists from Jewish organizations, (2) stratification of remaining non-organization list households by available measures of geographic incidence of the Jewish population, and (3) use of predictive modeling or vendor data to reach likely Jewish households not on Jewish organization lists.

Section 2 describes the sampling methodologies that NORC at the University of Chicago and the Cohen Center for Modern Jewish Studies at Brandeis University employed to conduct these studies. Section 3 describes data collection outcomes for these studies based on the designs. Section 4 summarizes conclusions and lessons learned regarding implementation of ABS for Jewish community studies.

2. Methodology

2.1 Overview

Sample designs for Jewish community studies face the challenges of providing accurate estimates about a rare population, providing estimates for small subgroups, and reaching Jewish households who are less engaged in Jewish organizations and activities. One method often employed for Jewish community studies is to sample from membership and participant lists from Jewish organizations. Without a design to sample outside of these lists, however, it would be very difficult to develop representative estimates of the community. Such designs do not reach households not involved with these organizations, and the characteristics of these households can, and typically do, differ from households involved in Jewish organizations across measures of demographics, Jewish practice, and affiliation.

To achieve the study goals, the sample designs employed for the 2020 Metropolitan Chicago and 2021 Greater Los Angeles studies sample from three main sources as depicted in Table 1. Details of the methodologies are provided in the technical appendices of Aronson et al. (2020) and Aronson et al. (2021).

First, we undertook a comprehensive process to utilize lists from Jewish organizations of their participants and members, using information on their names, mailing addresses, email addresses, phone numbers, and if available membership in different demographic groups. Although a common practice (Dutwin et al. 2013), our current application maximizes the utility of this practice by making every effort to uncover and secure every possible list from any community organization that has a meaningfully large share of Jewish members. We combined and deduplicated these lists to develop the organization list portion of the sampling frame. Most households from organization lists are Jewish, and thus reaching these households by a combination of mail, email, and telephone results in a low cost per completed survey. Households from Jewish organization lists also have a higher level of Jewish engagement than the rest of the Jewish population, although there is a range from low to high Jewish engagement.

Table 1: Jewish Community Study Sample Design Overview

Sampling Source	**Sampling Frame**	**Jewish Incidence**	**Cost per Complete**	**Jewish Engagement Level**
Deduplicated Jewish Organization Lists	Jewish organization participant addresses	Very High	Low	A Range
Remaining Households Predicted as Likely Jewish	USPS Computerized Delivery Sequence File	Moderate	Moderate	Lower
Remaining Households Not Predicted as Likely Jewish, Stratified by Geographic Measures of Jewish Incidence	USPS Computerized Delivery Sequence File	Low	High	Lower

As discussed, a comprehensive sampling methodology with strong population coverage must include households beyond Jewish organization lists. Thus, we sample from remaining households in the study area who are not on Jewish organization lists, using a version of the U.S. Postal Service (USPS) Computerized Delivery Sequence File licensed from a vendor. This portion of the sampling frame is segmented into two main groups. First, we use either predictive modeling (Dutwin 2020) or vendor data to identify households more likely to be Jewish and then employ higher sampling probabilities for these households. We also sample from among remaining households in the area, and we stratify the sample by available measures of geographic incidence. To assure high population coverage while maintaining a cost-effective sample design, we employ higher sampling probabilities for households identified as likely Jewish. Among the remaining households, we use higher sampling probabilities in geographic areas where available indicators show higher Jewish density.

Households predicted as likely Jewish have a moderate likelihood of being Jewish, lower than households from organization lists but much higher than the remainder of the population. Thus, predictive modeling and use of vendor data support reaching households not on Jewish organization lists at a moderate cost per complete. Using predictive modeling or vendor data also helps gain completes from households who are less engaged Jewishly. The remaining households have a low likelihood of being Jewish and a high cost per completed interview. Still, stratifying by geographic measures of Jewish incidence supports more cost-effective contact to these less engaged households while offering high population coverage.

We now describe the specific methodologies employed in further detail, first discussing sample development from Jewish organization memberships lists, followed by the development of the address-based sample.

2.2 Jewish Organization List Sample Development

For each of the Metropolitan Chicago and Greater Los Angeles studies, Brandeis and NORC undertook in-depth processes collaborating with the communities to collect organization lists from a range of organizations serving the Jewish community, including the respective Jewish Federations, synagogues, day schools, early childhood centers, camps, youth organizations, social service agencies, and organizations providing programming for specific community subgroups. The goals were to include as many Jewish households as possible in the organization frame and to represent a diverse cross-section of the community participating in various kinds of organizations. Brandeis and NORC worked closely with the communities to identify and reach out to organizations to request membership and participant lists, while taking careful measures to guarantee the confidentiality of household members on these lists. More than sixty organizations were included for the Greater Los Angeles study, and more than forty were included for the Metropolitan Chicago study.

For each study, NORC deduplicated the records among organizations using information available from the files. To detail the specific processes for the Greater Los Angeles study, NORC used a deterministic record linkage model to identify and remove households appearing on multiple lists. Likely business addresses were removed based upon information from a version of the U.S. Postal Service Computerized Delivery Sequence File (CDSF), and names, mailing addresses, and contact information were cleaned across all files. Records were considered matches if they shared two or more of the following characteristics: exact name; exact address; exact phone number; and exact email. Records were also considered matches if they shared fuzzy matched names or addresses along with exact matched values from the other contact information fields. Fuzzy matches were determined using the *stringdist* function from the *stringdist* package (van der Loo 2014). A pair of strings was considered a fuzzy match if their distance was less than 1 as calculated by the *optimal string alignment* algorithm, a modification of the Levenshtein distance that allows for transpositions of adjacent characters. Records that loosely matched, such as those that had matching names, unit numbers, and fuzzy matched emails, were deduplicated manually.

For both studies, we took great care in developing a stratified sample design from the organization list frame to ensure representation of different population subgroups and to support estimates for different geographic regions: ten for Metropolitan Chicago and six for Greater Los Angeles. Lists were reviewed to determine their association with specific population subgroups and different anticipated eligibility and participation rates from different lists. Then specific lists, and records in lists, were assigned to strata to support reaching data collection targets for key subgroups and to reflect assumptions on anticipated data collection outcomes for different lists.

Households were ultimately assigned to a sampling stratum based upon region and list source as described above. To assure an even distribution of the sample across the geographic area, the study team employed systematic sampling (Kish 1965) to sort the frame within these sampling strata by region, ZIP code, block, street name, and street number and then sampling at regular intervals within strata.

2.3 Address-Based Sample Development

Address-based sampling was conducted for both studies from a frame including all households in each area from a geocoded version of the U.S. Postal Service CDSF licensed from a vendor after removing households from the organization list frame. The implementation of an address-based sample is essential for high coverage of the Jewish community and to include community members less engaged with Jewish organizations. Goals for designing the address-based sample included assuring broad coverage of the area Jewish communities and maintaining costs by increasing the likelihood of reaching Jewish households. To achieve these goals, we designed these samples focusing on two strategies:

1. Classifying Census block groups by the estimated prevalence of Jewish households in the block group; and

2. Utilizing either a predictive model or vendor data to determine households' likelihood of having at least one Jewish adult.

First, NORC developed a measure related to Jewish density at the Census block group-level. The resulting measures were developed by combining multiple input block group-level measures. For the Greater Los Angeles study, we used (1) the percentage of households in the block group on the deduplicated organization list frame, (2) the percentage of households in the block group identified by a vendor as likely Jewish, and (3) the percentage of households with persons with likely Jewish surnames. For the Metropolitan Chicago study, we used (1) the percentage of households in the block group on the deduplicated organization list frame and (2) the percentage of households in the block group identified by a vendor data source as likely Jewish.

For both studies, a principal component analysis was conducted to combine the measures, and the first principal component was taken as a correlate of Jewish density. Block groups were ranked by the resulting Jewish density measure and then grouped into categories based on ranges of the measure. For the Metropolitan Chicago study, households were stratified based upon vendor data identification as a likely Jewish household or not combined with classification based on the Jewish density measure.

For the Greater Los Angeles study, NORC employed predictive modeling of Jewish households in the sample design. From the address-based frame, NORC selected a first-stage sample of records from the CDSF stratified by the block group measure of Jewish density. Households were matched with a large set of variables available from a vendor as well a block group-level data from the 2020 U.S. Census Bureau Planning Database (U.S. Census Bureau 2020). The team then used gradient boosted machines (Yuan 2022) to estimate for each adult from the household the likelihood of the household having at least one Jewish adult using more than 1,200 variables across the data sources. Specifically, adults in each household were assigned a propensity score for likelihood of being Jewish, and then households with an adult with a propensity score above a cutoff value were designated as likely Jewish. Among 171,223 households selected for the first-stage sample separate from Jewish organization lists, 4.0% were predicted as likely Jewish. The address-based sample was ultimately stratified based on the combination of block group Jewish density classification and classification as predicted Jewish or not, with higher sampling rates for households predicted as Jewish and in higher density block groups.

Due to the substantial number of cases needing to be sampled to reach Jewish households in the lowest density block groups, the study team did not sample households in the lowest Jewish density block groups that were not identified as being likely Jewish in either study. With the concentration of the population in specific geographic areas as well as the population coverage from organization lists and the use of predictive modeling or vendor data, the study team estimates that the sample design covered at least 96% of Jewish households for both studies.

As for the organization list sample design, systematic sampling was conducted for both studies to draw an evenly distributed sample across the geographic areas within combinations of classification as likely Jewish or not, block group Jewish density classification, and region.

3. Outcomes

Important goals for both studies were to support estimates for small subgroups of the Jewish community and to reach households less engaged in Jewish organizations and activities. Figures 1 and 2 show the number of completed interviews for key subgroups out of 3,877 survey completes for the Metropolitan Chicago study and out of 3,012 completes for the Greater Los Angeles study, respectively. Both studies succeeded with the goal of reaching small subgroups. Large numbers of completes were obtained for households with no Jewish denomination as well as for households classified by Brandeis as having minimally involved Jewish engagement. The studies also succeeded with reaching smaller subgroups in these Jewish communities, which in Metropolitan Chicago includes Russian-speaking Jews, LGBTQ Jews, Israelis, and Jews of Color; and in Greater Los Angeles includes Israelis, Persian Jews, and Russian-speaking Jews.

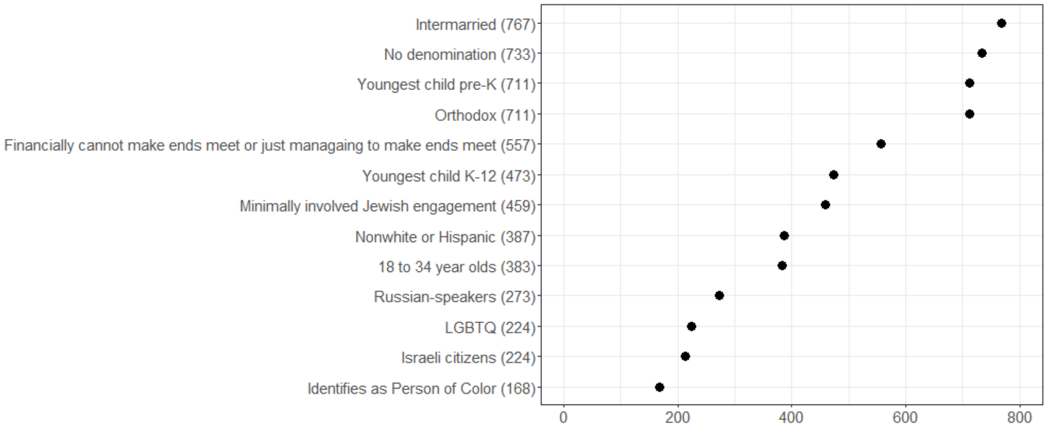

Figure 1: Summary of 2020 Metropolitan Chicago study completed interviews for key subgroups out of 3,877 total. Number of subgroup completes in parentheses.

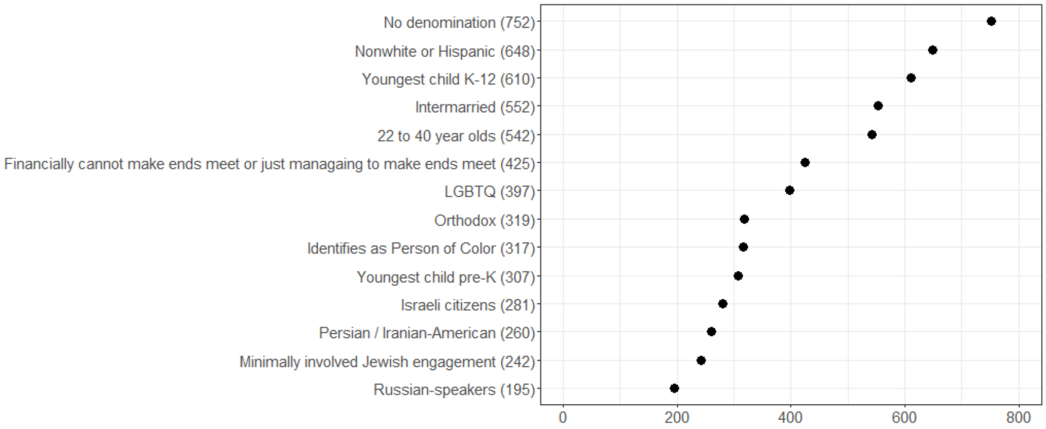

Figure 2: Summary of 2021 Greater Los Angeles study completed interviews for key subgroups out of 3,012 total. Number of subgroup completes in parentheses.

Additionally, the Metropolitan Chicago study sought to provide geographic detail, supporting estimates for ten total regions including three in the City of Chicago and seven outside of the City. The study succeeded achieving two hundred completes or more in nine of the regions, and only missing the target to support more detailed estimates in the region with the smallest Jewish population. Figure 3 shows a graphic from the Metropolitan Chicago Jewish Population Study (MCJPS) Interactive Mapping Tool (https://www.juf.org/Population-Study/) showing an example of the regional detail available for estimates.

A further goal for each study was to cost-effectively reach Jewish households while providing high coverage of the Jewish community in the area. NORC employed predictive modeling, use of vendor data, and geographic stratification to cost-effectively reach Jewish households. Figure 4 shows the eligibility rates for both studies for four different sample groups: (1) organization list sample, (2) predictive model or likely Jewish sample, (3) remainder sample in high density Jewish block groups, and (4) remainder sample in medium density Jewish block groups. In this graphic, the eligibility rate is defined as the percentage of households with determined eligibility status that were eligible for the interview—having at least one Jewish adult. The organization list sample had as expected by far the largest eligibility rates: 73.8% in Greater Los Angeles and 78.5% in Metropolitan Chicago. However, predictive modeling and vendor identification of likely Jewish households were able to provide eligibility rates much greater than that of the general population: 26.6% using predictive modeling in Greater Los Angeles and 22.1% when using vendor data in Metropolitan Chicago. Examining the remaining sample in geographic areas with higher Jewish density, the eligibility rates remained reasonable even though this sample excluded organization list and likely Jewish households, 8.7% in Metropolitan Chicago and 7.7% in Greater Los Angeles. Further, in the medium density Jewish block groups, the studies attained eligibility rates of 4.0% in Metropolitan Chicago and 4.7% in Greater Los Angeles.

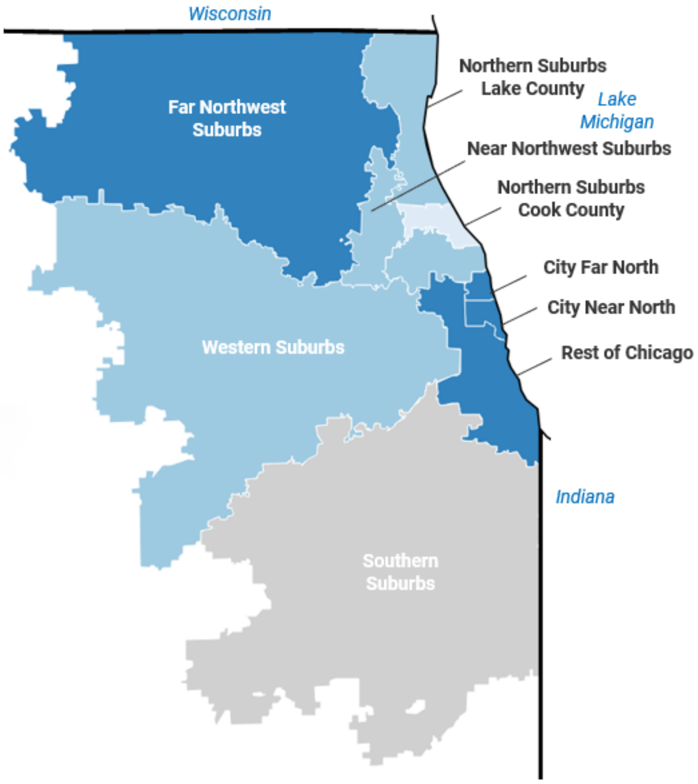

Figure 3: Graphic based on Metropolitan Chicago Jewish Population Study (MCJPS) Interactive Mapping Tool (https://www.juf.org/Population-Study/) showing regional detail available for estimates.

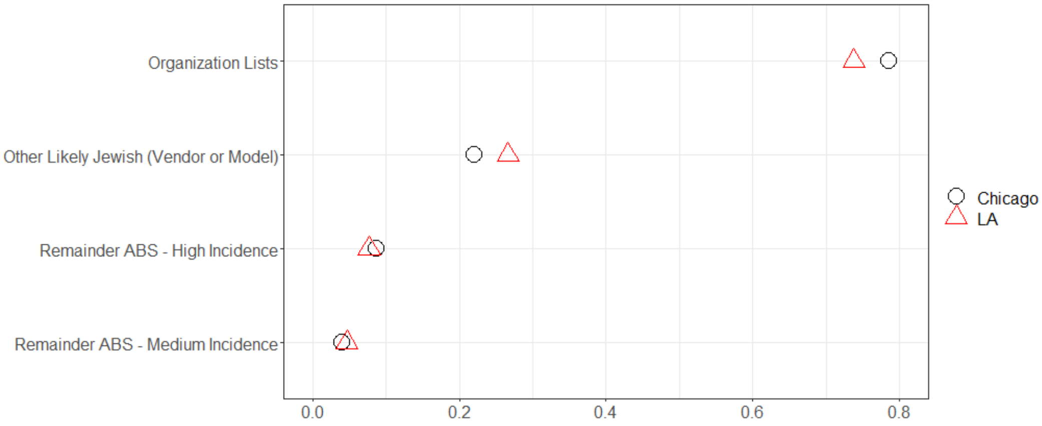

Figure 4: 2020 Metropolitan Chicago Study and 2021 Greater Los Angeles Study eligibility rates by sampling source.

4. Conclusion

Our experiences with the 2020 Metropolitan Chicago Jewish Population Study and the 2021 Study of Jewish Los Angeles demonstrate that address-based sampling can be adapted to conduct high quality studies of metropolitan area Jewish populations. Address-based sampling provides advantages of known probabilities of sampling for every household in the population and high population coverage.

We highlight the strategies that made these two studies successful and are critical for meeting the goals of supporting estimates for small subgroups, reaching less engaged members of Jewish community, and maintaining high population coverage at reasonable cost. First, we undertook great care and planning to develop organization list frames including membership and participation lists from a range of Jewish organization serving different segments of the Jewish community. We analyzed the information from lists to develop a sample designed to reach data collection targets for key subgroups. Second, for households not available from organization lists, we found predictive modeling in the Greater Los Angeles study to be highly effective in identifying Jewish households. Finally, we employed geographic stratification with available measures of Jewish density and used higher sampling rates for areas with higher Jewish incidence.

The strategies employed for these two studies are promising for the practice not only of Jewish community studies, but as well for other studies of rare populations. Utilizing many organizational lists for sampling is a detailed undertaking, but effective, although it may not be possible in all settings. Methods for address-based sampling utilizing predictive modeling and geographic stratification hold further promise to be adapted to support study of other rare populations. The lessons from conducting these surveys are generalizable to a range of applications and informative for good survey practice to study rare populations.

Acknowledgments

The authors thank our colleagues at NORC at the University of Chicago and Brandeis University for their important contributions to the sample design and methods employed for the 2020 Metropolitan Chicago and 2021 Greater Los Angeles Jewish community studies. From NORC, we thank Stephen Cohen, Ned English, Julie Banks, Evan Herring-Nathan, Chang Zhao, Patrick Coyle, Peter Herman, Katie Archambeau, Andrea Malpica, Becki Curtis, Frankie Duda, Shannon Nelson, Margrethe Montgomery, Steven Pedlow, Michael Yang, and Ed Mulrow. From Brandeis, we thank Janet Aronson, Matt Boxer, and Matthew Feinberg.

In addition, we thank our partners at the Jewish United Fund of Metropolitan Chicago and the Jewish Federation of Greater Los Angeles for sponsoring these studies and for their collaboration to support these study designs.

References

Aronson, J.K., Boxer. M., Saxe, L. (2016). *Contemporary Jewry* Vol. 36, No. 3, Special Issue: Jewish Community Studies (October 2016), pp. 361-380.

Aronson, J.K., Boxer, M., Saxe, L., Seeskin, Z. H., Cohen, S. (2021). 2020 Jewish Chicago: Who We Are: Population Study. https://www.jewishdatabank.org/databank/search-results/study/1127

Aronson, J.K., Brookner, M., Saxe, L., Bankier-Karp, A., Boxer, M., Seeskin, Z.H., Dutwin D. (2022). 2021 Study of Jewish LA. https://www.jewishdatabank.org/databank/search-results/study/1164

Dutwin, D., Ben Porath, E., and Miller, R. (2013). U.S. Jewish Population Surveys: Opportunities and Challenges. Studies of Contemporary Jewry, 17, 2013.

Dutwin, D. (2020). Feedback Loop: Using Surveys to Build and Assess Registration-Based Sample Religious Flags for Survey Research. *Big Data Meets Survey Science: A Collection of Innovative Methods*, 535-559.

Harter, R., Battaglia, M. P., Buskirk, T. D., Dillman, D. A., English, N., Fahimi, M., Frankel, M. R., Kennel, T., McMichael, J.P., McPhee, C. P., Montaquila, J., Yancey, T., & Zukerberg, A. L. (2016). Address-based sampling. Prepared for AAPOR Council by the Task Force on Address-based sampling, Operating Under the Auspices of the AAPOR Standards Committee. https://www.aapor.org/Education-Resources/Reports/Address-based-Sampling.aspx

Kish, L. (1965) *Survey sampling*. John Wiley and Sons, Inc., New York.

Lavrakas, P., Benson, G., Blumberg, S., Buskirk, T., Cervantes, I. F., Christian, L., Dutwin, D., Fahimi, M., Fienberg, H., Guterbock, T., Keeter, S., Kelly, J., Kennedy, C., Peytchev, A., Piekarski, L., & Shuttles, C. (2017). The future of U.S. general population telephone survey research. AAPOR Task Force Report.

Link, M. W., Battaglia, M. P., Frankel, M. R., Osborn, L., & Mokdad, A. H. (2008). A Comparison of Address-Based Sampling (ABS) Versus Random-Digit Dialing (RDD) for General Population Surveys. *Public Opinion Quarterly*, *72*(1), 6-27.

Pew Research Center. (2021). Jewish Americans in 2020. https://www.pewresearch.org/religion/2021/05/11/jewish-americans-in-2020/

Saxe, L., Tighe, E., and Boxer, M. (2014). Measuring the Size and Characteristics of American Jewry: A New Paradigm for an Ancient People." The Social Scientific Study of Jewry: Sources, Approaches, Debates. Ed. Uzi Rebhun. New York: Oxford University Press, 37-54.

U.S. Census Bureau. (2020). 2020 U.S. Census Bureau Planning Database. https://www.census.gov/topics/research/guidance/planning-databases.2020.html

van der Loo, M. P. (2014). The *stringdist* Package for Approximate String Matching. *R Journal*, *6*(1), 111.

Yuan, J. (2022). R Package *xgboost*: Extreme Gradient Boosting. https://cran.r-project.org/web/packages/xgboost/xgboost.pdf