Estimating Foreign Born Scientists and Engineers (FBSE)

Application of the Network Scale Up Method Using AdTech Data

November 2023

PREPARED BY:

Accenture Federal Services for the **National Science Foundation**

Amir Bagherpour, PhD, Data Scientist & Managing Director

Nolan Phillips, PhD, Senior Data Scientist

Heather Patsolic, PhD, Data Scientist

Alisha Kim, PhD, Data Scientist

Dan Gause, Data Scientist

Katie Hitchcock-Smith, Social Scientist & Analyst

Table of Contents

America's DataHub Consortium (ADC), a public-private partnership, implements research opportunities that support the strategic objectives of the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation (NSF). These results document research funded through ADC and are being shared to inform interested parties of ongoing activities and to encourage further discussion. Any opinions, findings, conclusions, or recommendations expressed above do not necessarily reflect the views of NCSES or NSF. Please send questions to ncsesweb@nsf.gov. This product has been reviewed for unauthorized disclosure of confidential information under NCSES-DRN24-044.

Key Findings

Analyses reveal that the Network Scale Up Method (NSUM) is a valid approach to estimating the number of foreign-born individuals obtaining science or engineering degrees. NSUM more efficiently estimates these numbers by leveraging other known sub-populations and how many individuals a respondent knows within those subpopulations. Using a corrective factor to adjust the ratio of US to foreign-born science and engineering degree holders, Facebook data estimate 1,686,226 and Amazon Mechanical Turk (AMT) data estimate 1,219,160 foreign-born scientists and engineers that are considering returning to their country of origin. These results indicate a potentially large "brain drain" for the US economy as these jobs are only increasing in demand. The AMT data appear to provide higher quality estimates compared to the Facebook data based on estimates of known sub-populations, which are treated as controls. Additionally, statistical tests find significant differences in the underlying populations from which these respondents draw their sub-population networks. Data were more cheaply obtained from AMT, at $2.73 per respondent, than when compared to Facebook, at $10.04. Both methods are significantly cheaper than a computer-assisted telephone interview, at $25 per respondent. Taken together, these results indicate that AMT is a cheaper, faster, and reliable method for obtaining sub-population estimates. However, it is worth noting that estimating other sub-populations that are less well known to AMT responders may not yield as accurate of estimates due to lack of familiarity. Future research should expand on these results and incorporate responder information to obtain more representative samples, which we believe could improve the results. The limits of the approach should be further assessed to better understand how many individuals are known in a sub-population to generate reliable estimates; this could be achieved using simulation studies. Substantively, these results confirm that there are a large portion of FBSEs in the United States, though more than a million are considering returning to their home country. In response, the US should consider pursing updated public policies to better assimilate these individuals and maintain their expertise, else risk losing their scientific skills and knowledge to other countries.

Abstract

The National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation (NSF) has conducted an online survey to estimate foreign born scientists and engineers (FBSE) who remain long-term in the United States. In addition to the numerical value of the data collected, the National Science Foundation is interested in the efficacy of the data collection methods, specifically, the network scale up method (NSUM) survey being implemented online. NSUM is an established statistical approach designed to avoid sampling bias and response bias when social desirability is a concern or access to a target population is challenging. The NSUM method is applied in this research to estimate the FBSE population on two different platforms, Amazon Mechanical Turk (AMT) and Facebook Advertising Technology (Ad-Tech). The successful application of NSUM, derived from online surveys, can open the door for more research by the National Science Foundation (NSF) and other organizations seeking alternatives to traditional stratified sampling approaches for hard-to-reach populations. In turn, advancements in advertising technology and distributed online communication technologies can reduce administrative costs while increasing convenience of reaching appropriate sample populations more efficiently.

Background

Digital Data Collection Methods

Social scientists and researchers are increasingly turning to digital methods to research demographic data among Americans. Much of the digital data from social media platforms being analyzed are taken from the metadata of social media users (Hargittai, 2021). While the metadata of social media users can be a useful method of analyzing various participant data trends over the years, there have been some quandaries with its collection. This data is sold by to researchers and companies. The use of participant data for commercial purposes is complicated due to legal, ethical, and practical concerns (Dewey Data, 2023). In recent years, there have been calls by users to have greater data protection and privacy measures of their metadata and digital footprint. In 2019 spurred on by Facebook's Cambridge Analytica controversy, members of Congress introduced the CONSENT Act (S. 2639), designed to protect the data of online users. Multiple other

legislative acts have been introduced on both a state and federal level, with some pending before the subcommittee on Digital Commerce and Consumer Protection to address online privacy (Loftsgordon, 2022). The majority of Americans say they are concerned, lack control, and have a limited understanding about how the data collected about them is used, while also favoring more regulation to protect personal information (McClain, Faverio, Anderson, & Park, 2023). For researchers, metadata collection may not be as useful as traditional research methods because it lacks the ability of social scientists to ask direct questions to participants and gain their informed consent to use their data (Cesare, Lee, McCormick, Spiro, & Zagheni, 2018).

Alternatively, the use of advertising technology such as Amazons Mechanical Turk and Facebook's Ad-Tech features are relatively new, but highly valuable resource for demographic research. Advertising technology refers to the programs, platforms, networks, and other tools publishers, advertisers, or others use to purchase, sell, and handle digital advertising (Ranne, 2023). Online advertising started in 1994 when a web magazine sold digital space to AT&T on their website. Since then, with the rise of digital media and platforms, online advertising has become ubiquitous in today's modern society and is estimated to reach $786.2 billion dollars by 2026 (Global News Wire, 2022).

The two platforms used in this study, Amazon and Facebook, both have wide-ranging benefits to reaching and sampling target demographics through their advertising technology. Amazon Mechanical Turk is a crowdsourcing platform that can reach a large population sample relative to other platforms (Bernard, Hallett, & Iovita, 2010; Chandler et al, 2018). In addition to online crowdsourcing platforms, social media platforms such as Facebook provide an alternate source for non-probability sampling. Facebook is one of the most widely used social media platforms in the US with demographic distributions similar to the national population (Auxier & Anderson, 2022). Facebook boasts 243.5 million users in the United States, almost 70% of the country's total population (Dixon, 2023). Beyond use capturing use metadata, advertising technologies also enables direct questions to participants. This serves as a powerful and transparent capability for surveying populations. The ability of online users to "opt-in" surveys also resolves ethical quandaries regarding consent when

compared to the practice of collecting metadata (Cesare, Hedwig, McCormick, Spiro, & Zagheni, 2018).

Network Scale-Up Method (NSUM)

In our research, the use of advertising technology allows surveyors to analyze the digital metrics of specific demographic statistics through using the Network Scale-Up Method (NSUM) of data collection. The NSUM is an established statistical sampling approach designed to avoid sampling bias and response bias when social desirability is a concern or access to a target population is challenging. The NSUM was first proposed by Bernard et al. (1989, 1991) following the 1985 Mexico City earthquake. The NSUM enabled researchers to estimate the number of people that died in the earthquake based on respondents' knowledge about their social contacts. In addition, the NSUM has been effectively used to estimate the size of populations at higher risk of HIV (Johnsen, Bernard, & Killworth, 1995) as well as other public-health relevant populations (Bernard, Hallett, & Iovita, 2010) and the estimation of the size of homeless populations and unreported crime (Killworth, McCarty, & Bernard, 1998).

The general format of an NSUM survey is “How many X do you know?”, also known as aggregated relational data. The wording of the questions is intended to distance the respondent from different subpopulations to accurately collect data on target populations that may have traditionally been difficult to collect from. In similar terms, the NSUM is premised on the broad probability calculation that for all individuals, the likelihood of knowing someone in a given subpopulation is derived from the size of the subpopulation divided by the overall population size (Maltiel, Raftery, & McCormick, 2016). For example, if a respondent knows 100 people total, and knows three people with HIV, then we can infer that 3% of the total population has HIV.

However, this is only true if the population respondent knows is representational to the population that is trying to be measured. The degree or personal network size of the respondent needs to be assessed, which can be accomplished by asking other questions related to the number of contacts in multiple subpopulations which are already known by researchers. For example, this study also asked “How many people do you know that currently reside in the US who… - Has the first name Mary?” or “How many people do

you know that currently reside in the US who… - Are a U.S. veteran?". These questions assume that a respondent should know roughly their degree multiplied by the proportion of people in each subpopulation. This information is used to "scale-up" the respondents answers to proportionally match the amount of people in a subpopulation.

Despite ease of use, the NSUM method contain some bias referred to as "barrier effects" (Killworth, 2003) (Killworth, 2006) (McCormick, Salganik, & Zheng, 2010). Barrier effects include the difficulty of taking into account the propensities of people knowing others in different groups or accurately estimating the control data. The NSUM method can also suffer from transmission bias, referring to a respondent not being aware that someone that they know is in a subpopulation, which is especially prevalent among groups which are stigmatized. Other biases include "recall bias" which refers to people to forget people that they know, or to overestimate the number of people they remember (Killworth, 2006).

Typically, NSUM is conducted as interviewer administered surveys, which is time consuming and resource intensive. Given the rise of online convenience samples, the effective use of an online-delivered NSUM remains an open research question, which is addressed by this research paper on foreign-born scientists and engineers.

Foreign-Born Scientists and Engineers (FBSE)

Immigration has had a long history in this country and has solidified the United States as a melting pot of culture and innovation. Among those immigrants are highly skilled and coveted scientists and engineers which have been crucial in key moments of United States history. Notable foreign-born scientists and engineers include Nobel Prize winning physicist Albert Einstein, inventor Nikola Tesla, Manhattan project theoretical physicist Maria Goeppert Mayer, and the cofounder of Google Sergey Brin (Long, 2017). Recognizing the importance of this demographic, the United States has implemented key policy decisions and legislation to allow easier emigration for scientists and engineers to North America from post-World War II Europe, typically through H-1B visas. Since then, thousands of foreign-born scientists and engineers have emigrated to America to use their skills and experience to benefit the American economy and inspire scientific innovation.

Foreign-born immigrant is a broad category, ranging from naturalized citizens and long-term U.S. residents to recent immigrants who compete in U.S. job markets and whose main social, educational, and economic ties are in their countries of origin (National Science Foundation, 2021). Science, technology, engineering, and math skills are often highly transferable across borders and are applicable to fast growing science and technology fields. Countries of origin are also known to benefit from emigration, with more capital being sent back to improve living standards from higher incomed individuals (Migration Policy Institute, 2022).

In 2015, foreign-born persons accounted for 29% to 30% of college educated workers employed in the science and engineering fields. Foreign-born scientists and engineers tend to have higher levels of education when compared to US born individuals. Among those in Science and Engineering workforce, 17% of foreign-born staff have a doctorate compared to U.S. native born individuals, with most citing the economic opportunities by moving and working in the United States for their motivation for applying for a visa (National Science Foundation, 2018).

"As more countries offer their students reasons to stay in their own country for their education or to return home after earning a degree, the U.S. could face a shortage in a critical segment of its workforce" (National Science Board, 2020). Now more than ever, as US native born individuals are having children later and less often, there will be fewer 18-year-olds entering college by 2026 and eventually the workforce as a highly educated and skilled conglomerate (Khanna, 2022). Through comprehensive surveys, understanding the future of a key demographic workforce such as foreign-born scientists and engineers can help provide insight in how to retain this talent and supplement the gap in the US labor force. To address this concern, the National Center for Science and Engineering Statistics (NCSES) intends to build evidence to understand the availability and demand for global science and engineering training and talent.

Methodological Approach

The NSUM survey was conducted online, with one set of participants recruited from Amazon Mechanical Turk (AMT) and another set of participants recruited from

Facebook over the course of five months. There were three main research questions to be considered when designing the survey:

1. Does an online non-probability sample combined with the Network Scale-Up Method (NSUM) allow the derivation of reliable, unbiased estimates of the foreign-born scientist and engineer population in the U.S.?
2. Does the use of an online NSUM improve cost and reduce public burden compared to existing methods without degrading quality?
3. Do response rates and time to complete survey data collection vary across platforms and compensation levels (e.g., AMT versus Facebook versus traditional methods)?

Samples obtained from online NSUMs were studied to examine the internal validity of NSUM questions as delivered on different online platforms. The results from the survey are compared with estimates from federal surveys and administrative data to provide insight on whether an online convenience sample NSUM delivered via AMT or Facebook can shed light on hard-to-count populations, specifically FBSE that reside in the US.

In practice, the NSUM questionnaire consists of twelve response questions and eighteen control questions grouped into four different categories: occupations, health-related, crime-related, and names. The control questions consist of known; countable populations used to estimate the fraction of the total population known to the respondent. As stated earlier, the size of the countable populations should be approximately 1-5% of the total population.

The survey is delivered only to AMT participants or Facebook users who reside in the United States, which is a delivery setting offered by the respective platforms. Specific to Facebook, survey delivery will be set to reach a diverse set of US users on the platform to ensure a random sample of delivery to US based platform users. Once participants are recruited, they are given a unique link to the online survey instrument, which is hosted specifically for this research allowing a single response per user. The platform also restricts duplicate responses.

Description of Samples:

Table 1: Descriptive statistics of respondents to education degrees obtained

Questions	Location	Mean	Standard Deviation	Median	Min.	Max.	Range
Facebook							
Biology or Life Sciences	US Born	9.362	10.331	5	0	30	30
	Foreign Born	4.351	7.717	1	0	30	30
Math or Computer Sciences	US Born	8.473	9.083	5	0	30	30
	Foreign Born	4.402	7.504	1	0	30	30
Physics or Chemistry	US Born	6.735	8.747	3	0	30	30
	Foreign Born	3.292	6.423	1	0	30	30
Social Sciences	US Born	9.783	10.361	5	0	30	30
	Foreign Born	3.042	6.001	1	0	30	30
Engineer	US Born	8.967	9.921	5	0	30	30
	Foreign Born	4.144	7.602	1	0	30	30
Science or Engineering	US Born	12.526	11.221	9	0	30	30
	Foreign Born	2.182	5.442	0	0	30	30
Amazon Mechanical Turk							
Biology or Life Sciences	US Born	4.241	26.806	1	0	670	670
	Foreign Born	1.211	3.827	0	0	30	30
Math or Computer Sciences	US Born	4.202	5.660	2	0	30	30
	Foreign Born	1.540	3.777	0	0	30	30
Physics or Chemistry	US Born	1.983	4.353	1	0	30	30
	Foreign Born	0.887	3.198	0	0	30	30
Social Sciences	US Born	4.906	6.981	2	0	30	30
	Foreign Born	1.137	3.526	0	0	30	30
Engineer	US Born	3.779	6.304	2	0	30	30
	Foreign Born	1.385	3.857	0	0	30	30
Science or Engineering	US Born	4.903	7.034	2	0	30	30
	Foreign Born	0.738	2.966	0	0	30	30

See Table 2 for descriptive statistics for the control questions.

To assess if the two samples draw from similar populations, we applied the Kalgomorov-Smirnov (KS) and Wilcoxon two-sample tests. The former quantifies the distance between the empirical distribution functions of the two samples; the latter compares the signed ranks of the two sample. Our team computed both since the KS test can be sensitive with larger populations. The tests assess the overall shape of the distribution and are less sensitive to outliers than other methods. In all cases, both found significant differences in the underlying distributions, except for the KS test on number of known people arrested, suggesting that the two groups are drawing from different sub-populations, as seen in Table B in the Appendix.

Key Findings & Results

As previously stated, the NSUM questionnaire consists of twelve response questions and eighteen control questions grouped into four different categories: occupations, health-related, crime-related, and names. For the latest version of the model, 16 of the 18 control questions were used, improving performance as compared with previous iterations of the model. As illustrated in Figure 1, our model using the Amazon Mechanical Turk (AMT) data is successful, defined as being within the estimation bounds, for five of the six characteristics in the names section, with the other name (slightly underestimated) and two of the health questions regarding stroke in the last three years (slightly overestimated) and cancer diagnoses in the past year (slightly underestimated) being close to within the estimate bounds. For the AMT data, the model gets progressively worse at estimating subpopulation counts that could potentially have higher levels of overdispersion. The model may benefit by removing two more of the common categories (e.g., number of people who smoke cigarettes and US veterans).

Using the same model parameters as AMT, we ran an analogous model using the Facebook data for comparison of the two survey techniques, with results also in Figure 1 below. This model was successful in estimating two of the 16 characteristics, the number of people who gave birth in the last year and the number of people whose home has been illegally entered by another without permission, both of which are well-underestimated in the AMT model. The Facebook model largely overestimated nearly every other subpopulation (12 of the 16), with no other category being within 1.35 times

overestimated or underestimated. The AMT model had half of the subpopulations within this range. This model may also benefit from removing the two more common categories, and may also benefit from more fine-tuning and possibly different choices of the scaling parameters than those used for AMT.

Figure 1: Model performance of control subpopulations

In comparing the two models, our team observed the AMT model maintaining higher accuracy with a tendency towards underestimation (10 of 16 control subgroups) and the Facebook model tending towards wider ranges of deviations from the literature-based values and overestimation (12 of 16 control subgroups). Furthermore, the Mean Absolute Error (MAE) of the AMT model is less than the MAE of the Facebook model, providing more confidence in the accuracy of the AMT model in predicting the unknown

FBSE subpopulations. Additionally, a trend towards overestimating the rarer populations in the Facebook model isn't as prominent with the AMT model. For example, two of the control groups account for more than 4% of the population each, and both subgroups were underestimated by both models. Thus, both models might benefit from the removal of these populations in future iterations. However, the remaining 14 subpopulations representing fewer than 2% of the population were relatively balanced in terms of under- overestimation in the AMT model, while all but two of them were consistently overestimated in the Facebook model.

Table 2: Mean estimate, Correction Factor (CF, mean/literature-based value), and Standard Error (SE) for Facebook as compared with AMT models for each of the known subpopulations.

		Facebook			AMTurk		
Category	**Literature**	**Mean**	**CF**	**SE**	**Mean**	**CF**	**SE**
Registered Nurse	3,072,700	6,834,137	2.22	5,608	5,459,168	1.78	201.9
School Teacher	3,971,816	9,774,354	2.46	3,239	7,178,753	1.81	280.3
US veteran	16,200,322	9,688,528	0.60	4,605	8,094,011	0.50	278.9
Smoke Cigarettes	28,300,000	6,413,674	0.23	5,810	7,500,259	0.27	386.2
Stroke in last 3 years	795,000	2,894,672	3.64	2,078	1,001,334	1.26	64.6
Diagnosed with cancer within last year	1,918,030	4,146,884	2.16	3,018	1,425096	0.74	65.4
Gave birth within last year	3,661,220	3,848,981	1.05	2,527	2,271188	0.62	113.7
Died within last year	3,273,705	5,156,371	1.58	2,022	1,676,044	0.51	56.3
Arrested within the last year	4,538,284	1,618,543	0.36	3,070	8,95,188	0.20	79.0
Had home illegally entered by another person without permission within last year	1,650,000	1,320,246	0.80	1,019	492,883	0.30	68.4
Mary	2,162,000	5,021,853	2.32	2,961	2,097,192	0.97	107.2
Elizabeth	1,612,390	4,619,018	2.86	2,787	1,806,977	1.12	66.8
Patricia	1,250,992	3,493,537	2.79	7,410	1,222,677	0.98	52.2
James	3,357,317	5,698,567	1.70	2,314	2,867,150	0.85	129.5
Robert	3,113,512	5,699,255	1.83	3,064	2,825,354	0.91	123.6
Michael	3,789,614	6,438,284	1.70	2,679	4,105,080	1.08	173.5

Table 3: Mean estimate, Correction Factor (CF, mean/literature-based value), and Standard Error (SE) for Facebook as compared with AMT models for each of the known subpopulations related to the unknown subpopulations and not used in the models. The correction factor is then applied to the estimates, partitioning the subpopulations based on the estimates for the unknown subpopulations.

		Facebook			**AMTurk**		
Category in US	**Total in US**	**Mean**	**CF**	**SE**	**mean**	**CF**	**SE**
Biological or Life Sciences	2,860,000	12,292,918	4.3	4,115	5,712,427	2.0	280
Math or Computer Science	2,786,608	12,456,943	4.5	4,734	7,532,251	2.7	315
Physics or Chemistry	2,300,000	9,805,474	4.3	3,829	3,746,475	1.6	202
Social Sciences	3,960,000	1,185,7344	3.0	4,225	7,864,915	2.0	355
Engineer	4,650,000	11,866,448	2.6	6,352	6,750,937	1.5	310
Science or Engineering	9,810,000	13,337,572	1.4	5,530	731,1127	0.7	313

Overall, the estimation of data from Amazon Mechanical Turk (AMT) from the latest model was decent, with the potential to be promising if the model is further iterated on. Currently, the AMT model is overestimating the population of scientists and engineers in the U.S., so it is probable the present model is also overestimating the amount of FBSE in the United States. However, taking the overestimation as a supplementary piece of information, in addition to the assumption that the US to non-US ratio of the total population of scientists and engineers is accurate, the overestimation can be repaired by a correctional factor which accounts for the known overestimation in both models. The correctional factor is calculated as the sum of US and non-US estimates (the top line model estimates in each degree category of Table 4) divided by the known total population of the respective category. Table 4 shows that the AMT model estimates, while predominantly overestimated, are closer than the Facebook model to the total numbers of scientists and engineers residing in the US, supporting the efficacy

of the FBSE estimates in the AMT model. The closest estimate in the AMT model was a modest underestimation of about 25% in the Science and Engineering category (corrective factor of .745). The Math or Computer Science category had the highest overestimation in the AMT model, about 170% above the known total population in that category. In comparison, all categories are overestimated in the Facebook model with the lowest (and closest) being an approximately 36% overestimation of the Science and Engineering category and the highest being an approximately 347% overestimation of the Math or Computer Science category. Notably, these categories match the lowest and highest extremes of the AMT model. However, the Facebook model has more variation in its overestimation of the scientific populations overall, leading towards higher trust in the AMT model, even after applying the correctional factor.

Table 4: Mean estimate, Correction Factor (CF, mean/literature-based value), and Standard Error (SE) for Facebook as compared with AMT models for each of the known subpopulations related to the unknown subpopulations and not used in the models. The correction factor is then applied to the estimates, partitioning the subpopulations based on the estimates for the unknown subpopulations.

		Facebook				AMTurk			
Degree	**Location**	**Mean**	**CF**	**Scaled**	**SE**	**Mean**	**CF**	**Scaled**	**SE**
Biological or Life Sciences	US Born	8,282,344	4.3	1,926,923	2,972.6	4,154,543	2	2,080,025	228.5
	Foreign Born	4,010,574	4.3	933,077	2,160.2	1,557,884	2	779,974	139.3
Math or Computer Science	US Born	8,306,115	4.47	1,858,071	3,544.2	5,480,792	2.7	2,027,657	252.9
	Foreign Born	4,150,828	4.47	928,536	2,438.7	2,051,459	2.7	758,951	161.1
Physics or Chemistry	US Born	6,655,793	4.26	1,561,202	3,194.9	2,597,915	1.63	1,594,887	156.1
	Foreign Born	3,149,682	4.26	738,798	1,461.5	1,148,560	1.63	705,113	124.8
Social Sciences	US Born	8,753,931	2.99	2,923,552	3,564.6	6,368,042	1.99	3,206,322	315.3
	Foreign Born	3,103,414	2.99	1,036,448	1,499.4	1,496,873	1.99	753,678	132.2
Engineer	US Born	8,033,016	2.55	3,147,827	4748.5	4,926,164	1.45	3,393,109	252.8
	Foreign Born	3,833,432	2.55	1,502,173	3,156.6	1,824,773	1.45	1,256,891	161.7
Science or Engineering	US Born	1,1044,998	1.36	8,123,774	4,248.5	640,2520	0.75	8,590,840	283.2
	Foreign Born	2,292,575	1.36	1,686,226	2,298.6	908,606	0.75	1,219,160	110

As stated previously, surveying demographics through utilizing advertising technology has been a relatively new frontier for social scientists. However, surveys proliferated through advertising technology have the potential to create an informed "digital census," reaching broad swaths of people in an ethical and pointed way, if done correctly. Several recent studies on utilizing Facebook's advertising platform for demographic research have been promising, including the sampling of hard-to-reach populations such as migrants or rural residents (Potzschke & Braun, 2017) (Rosenzweig, Bergquist, Hoffman Pham, Rampazzo, & Mildenberger, 2020). In a cross-national survey over seven European countries and the United States, researchers used Facebook's advertising platform to conduct social science research on the behaviors and attitudes in response to the COVID-19 pandemic. In the study, researchers compared self-reported and Facebook classified demographic information (sex, age, region of resident) with relatively accurate crossover to learn more about the effect of the COVID-19 pandemic (Grow, et al., 2022).

Compared to past methods of digital data collection, the utilization of advertising technology has multiple financial and practice benefits. When compared to costly established sampling techniques of random-route sampling, random-digit dialing, or face-to-face fieldwork, advertising technology merely requires buying advertising space on a social media platform. For example, in past attempts to measure the amount of foreign-born scientists and engineers, the National Science Foundation used a trimodal data collection approach: web survey, mail survey, and computer-assisted telephone interview (CATI), lasting over the course of seven months (National Center for Science and Engineering Statistics, 2021).

In this study, the National Science Foundation had two main costs, including paying the Amazon MTurk participants for their time ($4 per respondent) and the ads themselves at $8,880.69 total be shown over the course of two months (8/2/2023-10/9/2023) for Facebook data and one month for Amazon MTurk (8/1/2023-8/31/2023). For one month of promotion, AMT cost $1,757.40 and FB was $7,123.29 for two months, or $3561.65 per month. In total, 643 participant's data were used for Amazon MTurk and 790 for Facebook. One participant's data on FBSE was $2.73 for MTurk and $10.04 for Facebook. Both estimations are well below the cost per CATI interview of $25 for

household surveys reported and the $22.2 per phone interview which was reported in another study (Ballivian, 2021; Mahfoud, 2015).

Conclusion

The research illuminates new avenues and potential shortcoming for enumerating sub-populations. While Facebook appears to be a more expedient method for data acquisition, its quality compared to Mechanical Turk is lacking. The NSUM method, nonetheless, provides a more efficient method for estimating hard to count population, and online survey methods provide a fruitful avenue for researchers and practitioners to explore.

The specific findings vis-à-vis US and foreign-born scientists show that while most known scientists are US born, there is a sizable population, more than 900,000, foreign born scientists considering returning to their home country. US policy makers and higher education institutions need to consider better ways to persuade or incentive these individuals to remain in the American research ecosystem or else risk losing talented scientists trained in US institutions.

Appendix

Table A: Questions in survey

How many people do you know that currently reside in the US who were... - Born in the United States and has a college degree in biology or life sciences?
How many people do you know that currently reside in the US who were... - Born in the United States and has a college degree in math or computer science?
How many people do you know that currently reside in the US who were... - Born in the United States and has a college degree in physics or chemistry?
How many people do you know that currently reside in the US who were... - Born in the United States and has a college degree in social sciences?
How many people do you know that currently reside in the US who were... - Born in the United States and is an engineer?
How many people do you know that currently reside in the US who were... - Born in the United States with a degree in science or engineering?
How many people do you know that currently reside in the US who were... - Born outside the United States and has a college degree in biology or life sciences?
How many people do you know that currently reside in the US who were... - Born outside the United States and has a college degree in math or computer science?
How many people do you know that currently reside in the US who were... - Born outside the United States and has a college degree in physics or chemistry?
How many people do you know that currently reside in the US who were... - Born outside the United States and has a college degree in social sciences?
How many people do you know that currently reside in the US who were... - Born outside the United States and is an engineer?
How many people do you know that currently reside in the US who were... - Born outside the United States with a degree in science or engineering and are contemplating returning to their country of origin?
How many people do you know that currently reside in the US who... - Work as a Registered Nurse (RN)?
How many people do you know that currently reside in the US who... - Work as a school teacher?
How many people do you know that currently reside in the US who... - Are a U.S. veteran?

How many people do you know that currently reside in the US who… - Smoke cigarettes?
How many people do you know that currently reside in the US who… - Had a stroke in the last three years?
How many people do you know that currently reside in the US who… - Been diagnosed with cancer in the last year?
How many people do you know that currently reside in the US who… - Women who gave birth within the last year?
How many people do you know that currently reside in the US who… - Died within the last year?
How many people do you know that currently reside in the US who… - Died from COVID in the U.S.?
How many people do you know that currently reside in the US who… - Has been arrested within the last year?
How many people do you know that currently reside in the US who… - Has been physically attacked by someone within the last year?
How many people do you know that currently reside in the US who… - Had their home illegally entered by another person without permission within the last year?
How many people do you know that currently reside in the US who… - Has the first name Mary?
How many people do you know that currently reside in the US who… - Has the first name Elizabeth?
How many people do you know that currently reside in the US who… - Has the first name Patricia?
How many people do you know that currently reside in the US who… - Has the first name James?
How many people do you know that currently reside in the US who… - Has the first name Robert?

How many people do you know that currently reside in the US who… - Has the first name Michael?

Table B: Non-Education Questions, mean estimate, standard deviation, median, minimum, maximum and range.

Questions	Mean	Standard Deviation	Median	Min.	Max.	Range
Facebook						
Registered nurse	4.18323	5.849908	2	0	30	30
School teacher	5.484472	7.340088	2	0	30	30
US veteran	6.21118	6.961266	4	0	30	30
Smoke cigarettes	5.772939	6.884998	4	0	30	30
Stroke in the last three years	0.757387	2.37463	0	0	30	30
Diagnosed with cancer within the last year	1.079316	2.575028	0	0	30	30
Women who gave birth within the last year	1.748056	2.877343	1	0	30	30
Died within last year	1.682737	2.942671	1	0	30	30
Died from COVID in US	1.195956	2.718149	0	0	30	30
Arrested within last year	0.671851	2.201586	0	0	30	30
Physically attacked by someone within last year	0.485226	2.074399	0	0	30	30
Home illegally entered within last year	0.356143	1.863506	0	0	30	30
Mary	1.614308	3.020787	1	0	30	30
Elizabeth	1.391913	2.602191	1	0	30	30
Patricia	0.937792	2.024581	0	0	30	30
James	2.211509	3.284352	1	0	30	30
Robert	2.180404	3.267196	1	0	30	30
Michael	3.155521	4.048854	2	0	30	30
Amazon Mechanical Turk						
Registered nurse	4.18323	5.849908	2	0	30	30
School teacher	5.484472	7.340088	2	0	30	30
US veteran	6.21118	6.961266	4	0	30	30
Smoke cigarettes	5.772939	6.884998	4	0	30	30

Stroke in the last three years	0.757387	2.37463	0	0	30	30
Diagnosed with cancer within the last year	1.079316	2.575028	0	0	30	30
Women who gave birth within the last year	1.748056	2.877343	1	0	30	30
Died within last year	1.682737	2.942671	1	0	30	30
Died from COVID in US	1.195956	2.718149	0	0	30	30
Arrested within last year	0.671851	2.201586	0	0	30	30
Physically attacked by someone within last year	0.485226	2.074399	0	0	30	30
Home illegally entered within last year	0.356143	1.863506	0	0	30	30
Mary	1.614308	3.020787	1	0	30	30
Elizabeth	1.391913	2.602191	1	0	30	30
Patricia	0.937792	2.024581	0	0	30	30
James	2.211509	3.284352	1	0	30	30
Robert	2.180404	3.267196	1	0	30	30
Michael	3.155521	4.048854	2	0	30	30

References

Auxier, B., & Anderson, M. (2022, May 13). *Social media use in 2021*. Retrieved from Pew Research Center: Internet, Science, and Tech: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Ballivian, A. (2021). Using Mobile Phones for High-Frequency Data Collection' Mobile Research Methods—Opportunities and Challenges of Mobile Research Methodologies. *Ubiquity Press*. Retrieved from https://www.ubiquitypress.com/site/chapters/10.5334/bar.c/download/293/

Bernard, H., Hallett, T., & Iovita, A. (2010). Counting hard-to-count populations: the network scale-up method for public health. *Sexually Transmitted Infections, 86*, 11-15.

Cesare, N., Hedwig, L., McCormick, T., Spiro, E., & Zagheni, E. (2018). Promises and Pitfalls of Using Digital Traces for Demographic Research. *National Library of Medicine*. doi:10.1007/s13524-018-0715-2

Cesare, N., Lee, H., McCormick, T., Spiro, E., & Zagheni, E. (2018). Promises and Pitfalls of Using Digital Traces for Demographic Research. *National Library of Medicine*. doi:10.1007/s13524-018-0715-2

Chandler, J. (2018). A Feasibility Study of Using Mechanical Turk to Test Survey Questions. *Mathematica Policy Research Report for the National Center for Science and Engineering Statistics*.

Chandler, J., Sinclair, M., & Hudson, M. (2017). A Feasibility Study of Using Mechanical Turk to Test Survey Questions. *Mathematica Policy Research Report for the National Center for Science and Engineering Statistics*.

Dewey Data. (2023, April). *Acquiring Corporate Data for Academic Research*. Retrieved from https://www.deweydata.io/blog/acquiring-corporate-data-for-academic-research

Dixon, S. (2023). Facebook users in the United States 2018-2027. *Statista*, p. 1.

Evans, D. S. (2009). The Online Advertising Industry: Economics, Evolution, and Privacy. *The Journal of Economics Perspectives, 23*, 37-60.

Global News Wire. (2022). *Global Digital Advertising and Marketing Market to Reach $786.2 Billion by 2026 at a CAGR of 13.9%*. Retrieved from Research and Markets: https://www.globenewswire.com/en/news-release/2022/09/28/2524217/28124/en/Global-Digital-Advertising-and-Marketing-Market-to-Reach-786-2-Billion-by-2026-at-a-CAGR-of-13-9.html

Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., . . . Weber, I. (2022). Is Facebook's advertising data accurate enough for use in social science research? Insights from a cross-national online survey. *Royal Statistical Society*. doi:https://doi.org/10.1111/rssa.12948

Hargittai, E. (2021). *Research Exposed: How Empirical Social Science Gets Done in the Digital Age*. Columbia University Press. doi:https://doi.org/10.7312/harg18876

Johnsen, E., Bernard, H., & Killworth, P. (1995). A Social Network Approach to Corroborating the Numbers of AIDS/HIV+ Victims in the US. In *Social Networks* (pp. 169-187).

Khanna, P. (2022). *The Brain Drain That Is Killing America's Economy*. Retrieved from Time Magazine: https://time.com/6140707/americas-brain-drain-economy/

Killworth, P. (2003). Two Interpretations of Reports of Knowledge of Subpopulation Sizes. *Social Networks*, 141-160.

Killworth, P. (2006). Investigating the Variation of Personal Network Size Under Unknown Error Conditions. *Sociological Methods & Research*, 84-112.

Killworth, P., McCarty, C., & Bernard, H. (1998). Estimation of Seroprevalence, Rape and Homelessness in the U.S. Using a Social Network Approach. In *Evaluation Review* (pp. 289-308).

Loftsgordon, A. (2022). *Social Media and Your Privacy Rights*. Retrieved from Nolo: https://www.nolo.com/legal-encyclopedia/social-media-and-your-privacy-rights.html

Long, K. (2017). *20 Immigrants & Refugee Scientists Who Made America Greater (Part 1)*. Retrieved from StarTalk: https://startalkmedia.com/20-immigrants-refugee-scientists-who-made-america-greater-part-1/

Mahfoud, Z. (2015). *Cell phone and face-to-face interview responses in population-based surveys: how do they compare?*

Maltiel, R., Raftery, A., & McCormick, T. (2016). Estimating Population Size Using the Network Scale Up Method. *National Institute of Health*. doi:10.1214/15-AOAS827

McClain, C., Faverio, M., Anderson, M., & Park, E. (2023). *How Americans View Data Privacy*. Retrieved from Pew Research Center: https://www.pewresearch.org/internet/2023/10/18/views-of-data-privacy-risks-personal-data-and-digital-privacy-laws/

McCormick, T., Salganik, M., & Zheng, T. (2010). How many people do you know? Efficiently estimating personal network size. *Journal of the American Statistical Association* , 59-70.

Migration Policy Institute. (2022). *Global Remittances Guide*. Retrieved from Migration Policy: https://www.migrationpolicy.org/programs/data-hub/global-remittances-guide

Mullinix, K., Leeper, T., & Druckman, J. (2015). The Generalizability of Survey Experiments. *Journal of Experimental Political Science*, 109-138. doi:10.1017/XPS.2015.19

National Center for Science and Engineering Statistics. (2021). *National Survey of College Graduates*. Retrieved from https://www.nsf.gov/statistics/srvygrads-legacy/

National Science Foundation. (2018). *Immigration and the S&E Workforce*. Retrieved from Science & Engineering Indicators 2018: https://www.nsf.gov/statistics/2018/nsb20181/report/sections/science-and-engineering-labor-force/immigration-and-the-s-e-workforce

National Science Foundation. (2021). *Foreign-Born Workers as a Percentage of Individuals in Science and Engineering Occupations*. Retrieved from Science & Engineering State Indicators: https://ncses.nsf.gov/indicators/states/indicator/foreign-born-workers-to-se-occupations

Potzschke, S., & Braun, M. (2017). Migrant Sampling UsingFacebook Advertisements:A Case Study of Polish Migrantsin Four European Countries. *Sage Journals*. doi:https://doi.org/10.1177/0894439316666262

Ranne, K. (2023). *What is Advertising Technology – A Guide and Explanations*. Retrieved from Nexd: https://www.nexd.com/blog/advertising-technology-a-guide/

Rosenzweig, L., Bergquist, P., Hoffman Pham, K., Rampazzo, F., & Mildenberger, M. (2020). Survey sampling in the Global South using Facebook advertisements. *SocArXiv Papers*. doi:10.31235/osf.io/dka8f