LBNL-XXXX

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Assessment of Household Appliance Surveys Collected with Amazon Mechanical Turk

Hung-Chia Yang, Scott J. Young,
Jeffery B. Greenblatt, Louis-Benoit Desroches

Energy Analysis and Environmental Impacts Division
Energy Technologies Area

Sally M. Donovan

Consultant, Melbourne, Australia

May 2015

This work was supported by the U.S. Department of Energy under Lawrence Berkeley National Laboratory Contract No. DE-AC02-05CH11231.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, or The Regents of the University of California.

Ernest Orlando Lawrence Berkeley National Laboratory is an equal opportunity employer.

Assessment of household appliance surveys collected with Amazon Mechanical Turk

Hung-Chia Yang, Scott J. Young,
Jeffery B. Greenblatt, Louis-Benoit Desroches
Lawrence Berkeley National Laboratory, Berkeley, California, USA

Sally M. Donovan
Consultant, Melbourne, Australia

Abstract

Energy researchers need data on residential appliances to make effective recommendations for reducing energy consumption. For some products, however, traditional data sources do not have sufficient detail. Online surveys can provide a less expensive alternative for data collection, but the accuracy of these surveys is still unclear. Here, we compare the results of Amazon Mechanical Turk (AMT) online surveys of refrigerators, freezers, televisions, and ceiling fans to the nationwide Residential Energy Consumption Survey (RECS) deployed by the U.S. Energy Information Administration (EIA). To account for differences in demographic distributions between the online survey results and the general population, we weighted the results using standard cell weighting and raking techniques, as well as a combination of these, termed "hybrid." The weighted results gave a distribution of product ownership that was reasonably close to RECS, albeit with small, statistically significant differences in some cases. The cell weighting method provided a slightly better agreement with RECS than the other two approaches. We recommend AMT online surveys as an efficient and cost-effective way of gathering in-home use data on appliances that are not adequately covered by existing data sources.

Table of Contents

Abstract 1
1. Introduction 3
2. Methods 5
   2.1 Data Sources 5
   2.2 Amazon Mechanical Turk Survey Design 6
   2.3 Weighting Methodologies 8
      2.3.1 Cell weighting method 8
      2.3.2 Raking method 9
      2.3.3 Hybrid weighting method 9
   2.4 Comparison to Benchmarks 10
3. Results 11
4. Discussion 15
5. Acknowledgments 18
6. References 19
7. Appendices 22
   7.1 Difference in Phrasing Demographic Questions between AMT and RECS 22
   7.2 Determining Order of Demographic Variables for Cell Weighting 23
   7.3 Demographic Distribution Comparison for Refrigeration Products 26
   7.4 Product Ownership Questions 27

1. Introduction

Reliable and targeted information on household appliances is vital to the development of energy efficiency policies. The share of U.S. household energy used by appliances and consumer electronics has increased from 17% in 1978 to 35% in 2009 (U.S. DOE:EIA 2013a; McNary and Berry 2012). This number is expected to increase in the future with growing demand for more and better household appliances and consumer electronics. Energy researchers are striving to develop effective and economically viable strategies for reducing energy consumption. In order to do this, they require detailed data on the prevalence, characteristics and usage patterns of such devices in households. Researchers use several existing sources of household appliance information to perform cost-benefit analyses of energy efficiency. These include public government-produced data from the Residential Energy Consumption Survey (RECS) administered by Energy Information Administration (EIA) (U.S. DOE: EIA 2012), the Federal Trade Commission (FTC)[1], ENERGY STAR administered by the Environmental Protection Agency (EPA)[2], and the California Energy Commission (CEC)[3]. Private market research firms such as NPD Group[4], The Nielsen Company[5], IMS Research[6], and DisplaySearch[7], can provide information on annual shipments and the number of devices in homes, but this information is expensive to purchase and often lacks detailed appliance usage in households. There are also a number of other sources of information, such as the Association of Home Appliance Manufacturers (AHAM)[8], Appliance Magazine[9], manufacturer websites, and retailer websites. These data can be used to estimate device sales, numbers and types of devices per household, unit energy consumption, and national energy use.

Despite the abundance of information that these sources provide, there remain critical data gaps, in some cases making estimates of device ownership and energy use very uncertain. Large public surveys cover a wide range of devices, but they do not always provide the details needed for calculations of energy use. For example, the latest RECS survey in 2009 (U.S. DOE:EIA 2012) indicates the number of refrigerators and freezers in a home, but only very coarse information on the capacity of these appliances. Detailed information on capacity, however, is crucial for developing economic models of energy use. RECS also has no information about refrigerator cooling technology (i.e., vapor compression, thermoelectric, or absorption cooling) or less common refrigeration products such as wine chillers. In addition, existing data sources contain little information on the interaction between two types of appliances, such as the combined use of a ceiling fan and air conditioner. This information is important when determining how

[1] http://www.ftc.gov/bcp/conline/edcams/eande/appliances/index.htm
[2] http://www.energystar.gov/
[3] http://www.energy.ca.gov/
[4] https://www.npd.com/wps/portal/npd/us/home/
[5] http://www.nielsen.com/us/en.html
[6] http://www.imsresearch.com/
[7] http://www.displaysearch.com/cps/rde/xchg/displaysearch/hs.xsl/index.asp
[8] http://www.aham.org/
[9] http://www.appliancemagazine.com/

changes in the use of one appliance will affect the energy consumption of a second appliance.

Field monitoring can be used to directly characterize appliance prevalence and energy use (e.g., Lanzisera et al. 2013; Greenblatt et al. 2013a; Mercier & Moorefield 2011). Budgetary constraints and project deadlines, however, limit the number of locations and the period of time over which data can be collected. For example, ceiling fans are likely to be used more frequently in warmer climates and during the summer. Capturing the seasonal variation in ceiling fan use for different climates would therefore require monitoring in multiple locations for at least one year, which may be cost-prohibitive. The use of whole-house high-resolution smart meter data can make it easier to remotely measure energy consumption from many households. It is difficult, however, to detect and analyze specific appliances from this data, as there are not yet viable methods to identify and disambiguate individual appliances from the whole-house energy trace (Zeifman & Roth 2011).

Recently, researchers in a number of fields have used internet-based surveys to acquire targeted information in a short time with reduced cost. Many survey firms maintain panels of participants who are willing to perform surveys (Couper 2000). Additionally, researchers have posted surveys on the classifieds website, Craigslist.com (Hicks & Theis 2013; Attari et al. 2010), and the Amazon.com crowd-sourcing marketplace, Amazon Mechanical Turk (Paolacci et al. 2010; Attari 2013). These internet-based survey techniques have become popular partially because non-response and the increase in cell phone usage have raised concerns about the coverage of random-digit-dialing telephone surveys (Baker et al., 2010). Compared to traditional telephone, mail and in-person surveys, online surveys can yield a higher response rate and reach many people in a short time with a relatively low cost (Goodman et al. 2013). Online surveys also eliminate the need for researchers to hand-code the survey responses when doing the survey analysis (Cobanoglu et al. 2001). As a result, this type of survey could allow energy researchers to perform multiple surveys, each targeted at the information and population they want, within time and budget constraints.

Although internet-based surveys are becoming more popular, the accuracy of their results is debated. Some studies have suggested that online surveys provide accuracy comparable to other methods (Gosling et al. 2004), but the main disadvantage of online surveys is the potential for selection (or sampling) bias, a systematic error due to the non-random sampling of a population. This arises because participants in online surveys usually volunteer to participate in the survey or survey pool, as opposed to being randomly chosen to participate (Baker et al., 2010). To account for this bias, a weighting procedure based on demographic characteristics is frequently applied to make the demographic distribution of the online respondents more similar to that of the general population (Kalton and Flores-Cervantes 2003). Weighting can correct some of the selection biases in a non-probability sample, but it might not remedy all of the biases (Baker et al. 2013).

Researchers in several domains have estimated the bias of non-probability internet surveys by comparing survey results to benchmark data, such as administrative records or a large probability-sample survey. These studies have found mean absolute biases ranging from 3 % of respondents for a civic engagement study (van Ryzin 2008), to 5.4 % of respondents for a study of health indicators (Bethell et al. 2004), and 4.8 to 8.9 % of respondents for a study that looked at a variety of administrative and behavioral indicators (Yeager et al. 2011). Studies have also shown that demographic weighting can remove up to approximately 60 % of the bias that exists before weighting (Tourangeau et al. 2013).

In order to determine the potential of internet surveys for collecting appliance usage data, the current paper measures the bias we found in several non-probability internet surveys of household appliances. To do this, we compare estimates of appliance ownership that we collected with surveys on Amazon Mechanical Turk, to benchmark estimates of appliance ownership reported in RECS 2009. We compare the proportions of households with specific numbers of appliances, and we compare three different demographic weighting methods we used in the surveys. These comparisons allow us to estimate the bias that can be expected for appliance surveys performed on Amazon Mechanical Turk, and to determine whether this form of online survey is a viable complement to other methods of data collection in the energy research field.

2. Methods

2.1 Data Sources

We used data from three Amazon Mechanical Turk (AMT) surveys of residential appliance use. AMT is an internet marketplace that allows requesters to post tasks and workers to complete those tasks in exchange for a monetary payment. Amazon started AMT as a way for programmers to crowd-source tasks that require human intelligence, but it has also become popular as way to perform surveys of the general population (Paolacci et al. 2010; Buhrmester et al. 2011). While AMT can serve as a marketplace for connecting participants with surveys, it is not a full-service survey firm. For example, demographic information must be obtained via explicit survey questions, and the ability to completely define the participant pool prior to the survey is limited. Several articles have reviewed the positives and negatives of using AMT for surveys and research (Paolacci et al. 2010; Buhrmester et al. 2011; Goodman et al. 2013). For the current study, we used data on refrigerators and freezers[10] from a survey of refrigeration products (Greenblatt et al. 2013b); we used data on televisions from a survey of televisions and set-top boxes[11]; and we used data on ceiling fans from a survey on ceiling fans (Kantner et al. 2013). Each of the surveys was performed for a separate project, and they focused

[10] "Refrigerator" here refers to a stand-alone refrigerator (sometimes called an "all-refrigerator") that has no freezer compartment, as well as a refrigerator-freezer that possesses both refrigerator and freezer compartments. "Freezer" here refers to a stand-alone freezer, as opposed to a freezer that is part of a refrigerator.

[11] Set-top boxes are devices that provide TVs with video content from a cable, satellite or internet service provider.

on information not available from other sources. As a result, appliance ownership was the only survey topic that we could readily compare to benchmark data.

RECS, administered every four years by the EIA, provided a benchmark for the AMT surveys in our analysis. RECS has been the main source of nationally-representative residential energy use information for many years, and it thoroughly covers the most common household products. RECS is based on multi-stage area probability sampling, starting with a random selection of counties, dividing these into census segments, and then randomly selecting from segments (U.S. DOE:EIA 2013b). The number of counties and segments is controlled so that average energy consumption is obtained at multiple levels including national, census region, census division, and individual states or groups of states within census divisions (U.S. DOE:EIA 2011). RECS includes hundreds of questions about appliance ownership and usage, home characteristics, and demographics. The most recent survey (RECS 2009) contains more than 12,000 household samples representing every Census region (McNary and Berry 2012). Each RECS household sample is assigned a weight, based on U.S. Census Bureau data, indicating the number of households it represents (U.S. DOE:EIA 2013b).

2.2 Amazon Mechanical Turk Survey Design

We posted each of the surveys as a task on the AMT marketplace. All surveys were restricted to participants at least 18 years of age, and U.S. residents. For the set-top box and ceiling fan surveys, participants were also restricted to those who had at least one set-top box and ceiling fan, respectively, in their primary residence. Within these restrictions, any AMT user was able to select to participate in the survey. The surveys were deployed until we received our desired number of responses, which was approximately 2,000 (see the first column of Table 1). We paid participants approximately $1.50 for completing the survey, although this amount varied slightly based on the number of questions. Studies have suggested that compensation rate affects mainly the data collection speed rather than the data quality (Buhrmester et al. 2011), and we set our fees based on a targeted hourly rate of $10-12 for median response-time workers to achieve moderate collection speed. We completed data collection for all surveys in less than three weeks, and the total survey costs were less than $5,000 each.

All of the AMT product surveys contained eight demographic questions that coincided with RECS 2009. The demographic questions were used as a basis for assigning weights to AMT survey data. The eight demographic variables were gender, zip code, Hispanic origin, race, education, household income, number of household members, and age of each household member. In some instances, the phrasing of the question varied slightly from the phrasing used in RECS, in order to be more appropriate for an online format. Details of these variations are outlined in section 7.1.

After comparing the initial round of responses to RECS, we found that some demographic categories were consistently over-represented across all surveys, while other categories were consistently under-represented. This phenomenon has been noted by other studies using AMT, including Paolacci et al. (2010) and Ipeirotis (2010). To

make the surveys more representative of the general population, we used a form of quota sampling to increase the number of respondents from under-represented demographics. For all three surveys, subgroup surveys were deployed for participants with no college education, households with people aged 60 and over, and participants who identified themselves as black/African-American. For ceiling fans, further subgroup surveys were deployed for one-person households, participants who identified themselves as Hispanic, and residents of the Mountain South (New Mexico, Nevada and Arizona), and West South Central (Arkansas, Louisiana, Oklahoma, and Texas) census divisions. The number of additional responses is shown in column 2 of Table 1.

In designing the surveys, we considered three common problems related to survey participants. These problems are described in Baker and Le Guin (2007) as "hyperactives," "fraudulents," and "inattentives." Hyperactives, also known as professional respondents, are people who use survey participation as a source of income, and often participate in many surveys and own membership to multiple survey platforms. Fraudulents deliberately give false information in order to be eligible to take part in the survey. AMT significantly reduces the potential risk of introducing poor quality responses from hyperactives and fraudulents, by requiring all participants to register with a credit card (Paolacci et al, 2010) and allowing survey posters to block workers from future surveys if they consistently give poor quality responses. In order to identify inattentives, we used two common approaches, as described in both Baker and Le Guin (2007) and Paolacci (2010). We included (1) one or more simple questions that any U.S. resident would know, such as "who is the current U.S. president?", and (2) a consistency check, by including two different questions that have the same answer. In this case we asked respondents to state the total number of household members, and then later asked them to identify the number of household members by age. The sum of household members across ages should equal the stated total number of household members in the earlier question, and if not, we considered the respondent failed the cheater question. Responses that had the wrong answer to one or more of these questions were omitted from the analysis. Responses with excessive skipped questions (in the range of three to ten questions depending on the survey) were also omitted. The numbers of responses received and final accepted responses for each survey are shown in Table 1. The majority of responses removed from the final analysis were inattentives.

Table 1: Details of AMT surveys used in the analysis

Survey product	General population responses received	Subgroup responses received	Total responses received	Final responses used in the analysis	Percentage of total responses that were used in the analysis
Refrigerators and freezers	2,330	1,100	3,430	3,021	88.1
Televisions	2,120	1,323	3,443	2,294	66.6
Ceiling fans	2,000	1,250	3,250	2,799	86.1

2.3 Weighting Methodologies

Applying weights is a common approach to making survey responses more representative of the general population. Kalton and Flores-Cervantes (2003) describe weighting as "a series of stages to compensate for unequal selection probabilities, nonresponse, non-coverage, and sampling fluctuations from known population values." While it was not possible to examine all possible weighting approaches, we applied three weighting methods to the AMT survey respondents using RECS as the reference population. First, we used cell weighting, which directly compared to the combinations of demographics in the RECS sample. Second, we used raking, which weighted based on the marginal distribution of demographics. Last, we used a hybrid method, which combined elements of cell weighting and raking in an attempt to overcome their limitations. All three weighting methods were applied to the three surveys shown in Table 1.

2.3.1 Cell weighting method

With the cell weighting method, we directly compared a subset of the respondent demographics to the corresponding demographics in RECS. We then applied a weight so that the sum of all survey responses with a given demographic combination equaled the sum of RECS responses with the same demographic combination. The method is described by Kalton and Flores-Cervantes (2003). For our surveys, we used five variables to apply cell weighting. Even though eight demographic variables were collected in each survey, due to the number of non-responses for certain variables, the resulting weights did not change appreciably with the use of more than five variables. In dealing with non-response, our approach differs from Kalton and Flores-Cervantes (2003). We simply reduced the number of demographics used to weight particular responses. In rare cases, this meant the weight of a response could be based on a single variable if the respondent only answered the first demographic variable used in the cell weighting method. The weights for other responses in the same demographic subgroup were adjusted to ensure the totals remained the same. Two forms of non-response occurred during the weighting process. In some cases, a response in an AMT survey had a combination of demographics that did not exist in RECS, meaning the response could not be assigned a weight using the process described. In other cases, a null response (such as "don't know" or "decline to state") was received in the AMT survey, or a question was left blank. In both of these instances the response would simply be assigned a weight based on fewer demographics. The approach to non-response meant the weights were very sensitive to the order of the demographic variables used in the weighting process. We determined the most appropriate order by weighting the data multiple times using different orders of demographics, and then comparing the results to the demographic distributions in RECS. The weighting that gave the smallest difference in demographic distribution from RECS used this order: Census division, number of occupants, race, number of 20-29 year olds and education level. A detailed explanation of the determination of variables is in section 7.2.

2.3.2 Raking method

Raking, a widely used and well-established technique for weighting sample populations to the general population was chosen as an alternative technique to cell weighting. Raking is an iterative proportional fitting procedure that operates on the marginal distribution of the variables, rather than the combined distribution used in cell weighting. This makes the technique less sensitive to non-response. Several papers describe raking in detail, including Battaglia et al. (2009) and Deville et al. (1993). Kalton and Flores-Cervantes (2003) compare cell weighting and raking and state that while there are circumstances in which the choice of weighting technique is obvious, in situations where the combined data distributions of the weighting variables are available, the choice between cell weighting and raking is less clear. Therefore applying both techniques and comparing the resulting data is an appropriate decision-making approach. Raking was applied to the variables in the order race, income, the existence of household members under the age of 20,[12] Hispanic origin, gender, Census region, education level, age distribution and number of people in the household, although varying the order was not found to have a significant impact on the resulting weights. When encountering a null response for a certain demographic variable, a weight of one was assigned to that response. The process was iterated until the weight of each response changed by less than a threshold amount set to 0.5%. The number of iterations required to reach the convergence threshold was always below 100.

2.3.3 Hybrid weighting method

While both raking and cell weighting gave a more similar distribution of demographics to RECS compared to the unweighted results, there exist limitations to each method.[13] The cell weighting method requires at least one demographic variable to have a complete set of survey responses; and an appreciable sample size for both survey and reference data (at least ~2,000 observations with five demographic variables to weight). For the raking method, the correlation between demographic variables is generally not considered, which may result in a "knock-on" effect when a survey respondent belongs to an under-represented category for one demographic and an over-represented category in another demographic. For example, low education (e.g., less than high school diploma) was under-represented in our surveys, whereas households with 20-29 year olds were over-represented. Therefore, we would want to apply a large weight to a respondent with low education, and a small weight to a respondent in a household with 20-29 year olds. However, when a survey respondent is both of low education and in a household with 20-29 year olds, weighting to correct one of these categories will make the other one less representative.

We therefore developed a hybrid method to take advantage of the strengths of both the raking and cell-weighting methods and complement each method's limitations. For the

[12] We created this variable to distinguish households with member(s) under the age of 20 from those without.

[13] An example of demographic distribution comparison across three weighting methods, non-weighted AMT and the RECS for the refrigeration product survey can be found in section 7.3.

hybrid method, we first applied the raking method to weight all the variables that were not going to be used in the cell weighting method. The cell weighting method was then applied to the remaining variables. In the case of non-response at the first demographic variable included in the cell weighting stage, the weight generated from the raking method was used. After the cell weighting method had been applied, the sets of weights were multiplied together. That value was then multiplied by the ratio of the number of responses represented by a given set of demographics in the cell weighting method to the total raking weight for those responses. This last adjustment ensured the final set of weights still summed to the same total. We hypothesized that the hybrid method would potentially mitigate the "knock-on" issue in the raking method by using cell weighting to include part of the correlations among key demographic variables. In addition, the hybrid method also improved the demographic distribution of the variables that were not originally included in the cell weighting method and improved the final weighting results by providing a pre-weighted sample for the cell weighting method to start with. In determining which variables to include in the cell weighting part and raking part of the hybrid weighting process, we weighted using four demographics in cell weighting and five in raking, then we weighted using five demographics in cell-weighting and four in raking. The results were compared with RECS, and the method that produced the most similar demographic distribution to RECS was chosen as the final weighting approach. We found that applying the cell weighting method to five variables, as described previously, then combining with the raking method applied to all other variables, minimized the difference from RECS demographics for the set-top box and ceiling fan surveys. For the refrigeration product survey, applying cell weighting to four demographics (not including education) gave a demographic distribution that had smaller differences from RECS.

2.4 Comparison to Benchmarks

To estimate the accuracy of the weighting methods, we examined the agreement in product ownership between the AMT and RECS data. We analyzed refrigerators and freezers from the refrigeration products survey, televisions from the set-top box survey, and ceiling fans from the ceiling fan survey. As the set-top box and ceiling fan surveys were restricted to participants that owned at least one of these products, the RECS data set used as their benchmark was also restricted to the owners of these products for the analysis. We performed the analysis on televisions and not set-top boxes, because the set-top box survey and other data sources (Nielsen 2012) indicated that product ownership of set-top boxes changed substantially between 2009 when RECS was performed, and 2012 when the AMT surveys were deployed.

Product ownership questions ask respondents about the number of a given product that they have in their home. These questions usually also specify that the product is in use. For example, RECS asked "How many refrigerators are plugged-in and turned on in your home?", and the AMT refrigeration products survey asked "How many refrigerators are plugged-in at your home right now?" A complete list of the product ownership questions used in the analysis can be seen in section 7.4. To compare results, we compared the percentage of households with a specified number of appliances in AMT, to the

percentage of households with the same number of appliances in RECS. For example, we compared the percentage of households with three refrigerators in AMT to the percentage of households with three refrigerators in RECS.

Before making comparisons, we adjusted the responses from AMT and RECS to the same scale. Both surveys reported ownership up to only a maximum number of appliances per household, but the maximum number used in the AMT surveys was often different than the maximum number used in RECS. We converted these scales to a maximum of "10 or more" for ceiling fans, and "5 or more" for all other appliances. As a result, all responses with a number of appliances equal to or greater than the new maximum were combined together. No households in RECS had more than 3 freezers, so we used zero percent as the ownership value of "4 freezers" and "5 or more freezers" for RECS.

We calculated the percentage point differences between the percentage of households owning each number of appliances in AMT and the corresponding percentages in RECS (e.g., if 22% of households in AMT owned three refrigerators, and 20% of households in RECS owned three refrigerators, the percentage point difference would be 2%). We then plotted each difference with a 95% confidence level based on the normal approximation to the binomial distribution (Steel et al. 1996). Since more than two responses were possible in each question, we used the Bonferroni correction (Bailey 1980; Cherry 1996) to create slightly larger, individual confidence intervals that result in a 95% confidence level across all levels within a group. These intervals provide a more conservative estimate of the uncertainty in the quantities than the individual 95% confidence intervals, and we refer to these intervals as Bonferroni-corrected confidence intervals.

To summarize the comparison results, we calculated the mean and maximum absolute difference from RECS for each appliance. We then ranked each weighting method, including the unweighted results, to identify the method that gave the distribution of product ownership closest to RECS for each product. Finally, we averaged the mean absolute difference, maximum absolute difference, and ranks across all products. These average values allowed us to identify which method gave the most accurate distribution of product ownership overall.

We also broke out the comparison of results by demographic group. These results were similar to the results using all respondents, so we have not included the results by demographic group here. More details about this comparison can be found in Greenblatt et al. (2013c).

3. Results

Figs 1 and 2 show individual results for all ownership categories. Fig. 1 shows the product ownership reported by RECS households for each product. The majority of the households own at least one refrigerator; however, 70 % of households do not own a freezer. Ownership generally decreases for larger numbers of appliances, with the exception of televisions, for which most television-owning households have at least two

televisions rather than a single television. Fig. 2 shows the percentage point differences between AMT surveys and RECS, for AMT data unweighted and weighted using the three methods, along with uncertainty estimates. The figure shows that uncertainty varies across the possible responses for each product. In most cases, the uncertainty estimates representing group-wise 95% confidence intervals overlap with zero.

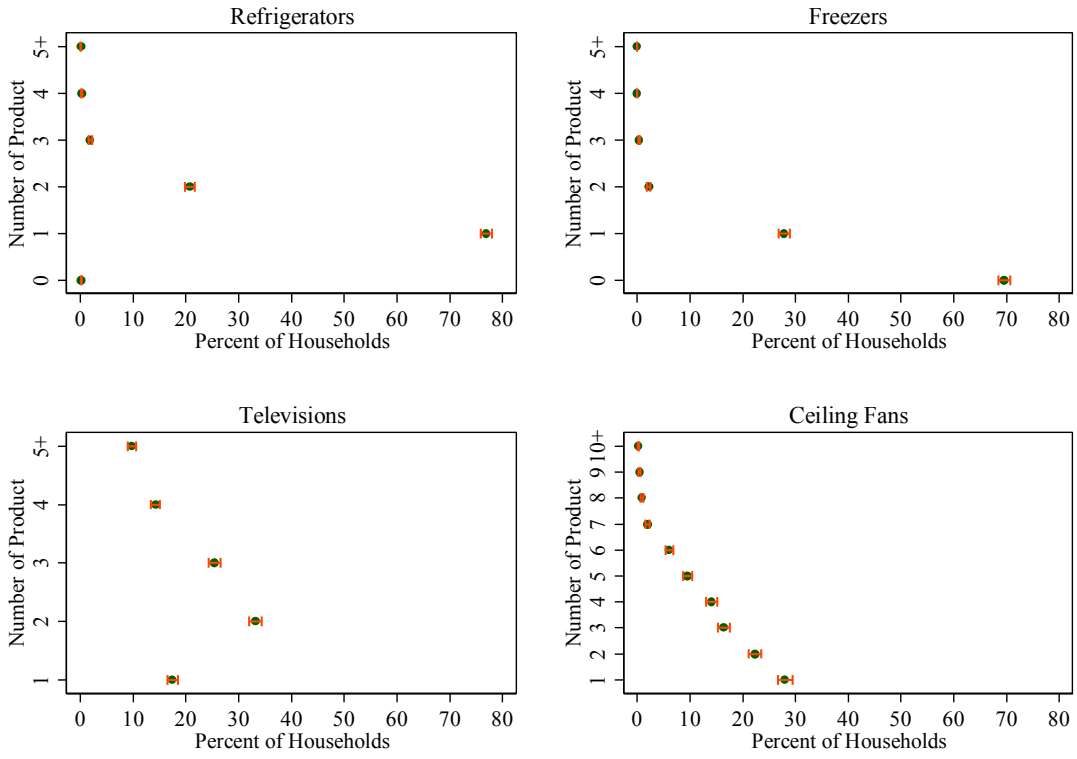


Fig. 1 Product ownership in RECS. Mean values are shown with dots, and uncertainty estimates represent Bonferroni-corrected 95% confidence intervals.

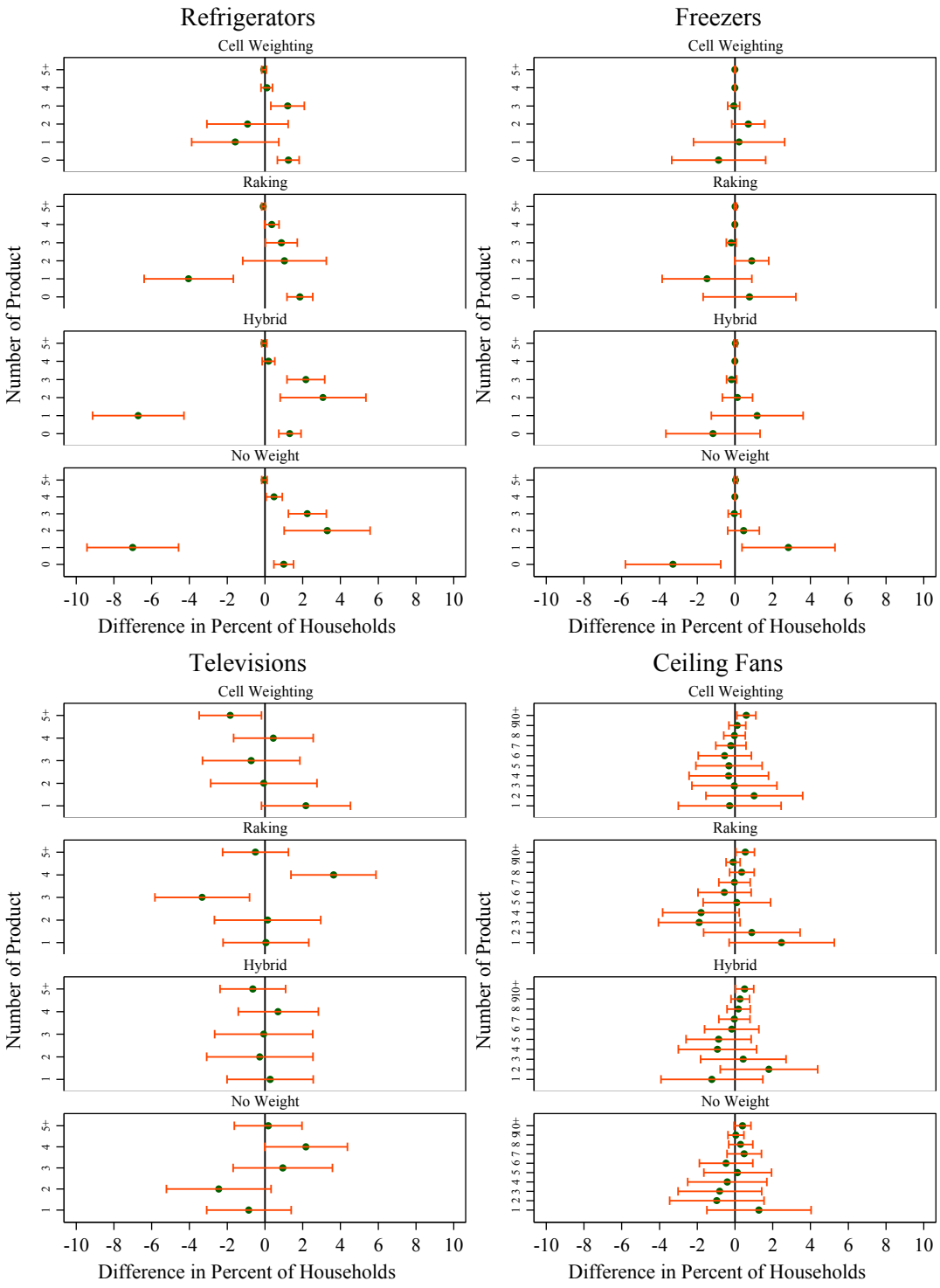


Fig. 2 Differences in percent ownership between AMT surveys and RECS reference samples for all four products and all weighting methods. Dots show the percentage-point difference, with a negative difference indicating that the value in the AMT survey was lower than the value in RECS. The uncertainty estimates represent Bonferroni-corrected 95% confidence intervals.

Table 2 summarizes the results of the comparisons, including the mean and the maximum absolute difference for each product and weighting method. These results quantitatively evaluate the improvement made by each weighting method. Cell weighting produced the best match to RECS for all products except televisions, for which the hybrid method provided the best agreement. Within the ranks, "1" indicates the best agreement with RECS (smallest absolute error) and "4" represents the worst agreement with RECS (greatest absolute error). By taking the average rank of each weighting method across all products, we found that cell weighting produces the best overall agreement with RECS, followed by the hybrid method. Raking produced the poorest match to RECS, and in some cases was ranked lower than the unweighted data. We also looked at other metrics, such as the mean squared error, but since the results are similar to the two metrics presented in Table 2, they are not shown here.

Table 2: Summary of differences in product ownership for all products

Evaluative Criteria	Refrigerators	Freezers	Televisions	Ceiling Fans	Average
Mean absolute error (% of participants)					
Cell Weighting	0.85	0.31	1.05	0.35	0.64
Raking	1.37	0.56	1.53	1.04	1.13
Hybrid	2.25	0.45	0.39	0.64	0.93
No Weight	2.34	1.11	1.32	0.53	1.33
Maximum absolute error (% of participants)					
Cell Weighting	1.57	0.87	2.17	1.02	1.41
Raking	4.04	1.48	3.63	3.69	3.21
Hybrid	6.71	1.18	0.70	1.80	2.60
No Weight	7.00	3.28	2.45	1.27	3.50
Rank of mean and maximum absolute error, relative to other methods					
Cell Weighting	1	1	2	1	1.25
Raking	2	3	4	4	3.25
Hybrid	3	2	1	3	2.25
No Weight	4	4	3	2	3.25

4. Discussion

In this study, we compared household appliance ownership estimated with AMT surveys, to the same information estimated with RECS 2009. Across four appliances, the mean unweighted absolute difference between surveys ranged from 0.5 to 2.3 % of respondents, and the maximum unweighted absolute difference between surveys ranged from 1.3 to 7.0 % of respondents. Across all appliances, the average unweighted absolute difference between AMT and RECS was 1.3 % of respondents. On average, all of the demographic weighting methods we used reduced the differences between AMT and RECS results. Cell weighting reduced the mean and maximum absolute difference the most, resulting in an average absolute difference between AMT and RECS across all appliances of 0.6 % of respondents. These results provide evidence of the bias to be expected from non-probability internet surveys of appliance usage, as well as indicating how much demographic weighting may reduce the bias.

The results of the current study suggest that the AMT appliance surveys, even without any weighting applied, can provide results on appliance ownership that are very similar to the results from RECS. The mean absolute bias of approximately one percent is small enough for the results to be useful for estimates of appliance energy use. The mean absolute unweighted bias we found in the current study is also smaller than the mean absolute unweighted bias of 3 to 6 % seen in three other studies of internet surveys (Bethell et al. 2004; van Ryzin 2008; Yeager et al. 2011)[14]. There are several possible reasons why our appliance surveys might have had smaller mean absolute bias than the surveys considered in the other studies. First, the AMT appliance surveys had more respondents than many of surveys considered in the other studies. Second, we used quota sampling to increase the number of responses from under-represented groups, whereas many of the other surveys did not. Third, the topics of the other surveys might be more susceptible to fluctuations than appliance ownership. For example, Yeager et al. (2011) and Bethell et al. (2004) asked about health, and Van Ryzin (2008) asked about political opinions.

This study also suggests that demographic weighting can reduce the bias of AMT appliance surveys. In our surveys, cell weighting reduced the mean absolute difference by approximately 50% of the unweighted difference. This value is close to the 60% upper bound of effectiveness found by Tourangeau et al. (2013). The raking and hybrid methods reduced bias for some surveys, but these methods also increased the bias in some cases (e.g. television and ceiling fan surveys). This inconsistent effect of weighting is similar to observations from other studies (Tourangeau et al. 2013; Yeager et al. 2011; Loosveldt and Sonck 2008).

[14] We calculated the range quoted in this sentence from numbers provided in the tables of the studies. For Bethell et al. (2004), the mean absolute difference between the online and benchmark values in the overall columns of Table 3 is 5.4%. For van Ryzin (2008), the mean absolute value of the 20 differences plotted in Figure 2 is 3%. For Yeager et al. (2011), the mean average percentage point absolute error, for the secondary and non-demographics, without post-stratification, of the non-probability sample internet surveys in Table 2 is 5.86%.

The raking method gave the poorest agreement of the three weighting methods for all products, and was worse than the unweighted results for televisions and ceiling fans. These results suggest two things. First, the poor performance suggests that correlations among demographic variables are likely important in estimating product ownership. These correlations were included in the cell weighting and hybrid methods, but not in the raking method. Second, the variability between products suggests that the relationship between demographic variables and product ownership could be product-specific. This issue is discussed in Battaglia et al. (2009), which states that weighting should focus on variables that exhibit strong associations with key survey outcome variables or are strongly related to non-response or non-coverage.

Although the cell weighting method was found to give the best performance in our analysis, we do not believe it will be the best method for all surveys. The choice of weighting method should be decided on by considering correlations between demographic variables as well as the correlations between demographic variables and the information the survey is trying to obtain. Further to this, it is not possible to apply the cell weighting method to all data sets, as described in Kalton and Flores-Cervantes (2003). It requires the survey data to have an appreciable sample size (~2,000 or more responses), have at least one complete set of responses, and have demographic combinations of variables in the benchmark sample that are the same as in the survey data. If the survey sample fails to satisfy these requirements, the cell weighting method would not generate robust inference of the survey data. Raking, on the other hand, only requires the marginal distribution of the variables and may be more appropriate to use for smaller sample sizes, higher levels of non-response or limited benchmark data. Researchers must therefore consider all the aspects of their own survey and then decide on the most appropriate weighting method.

There are at least four limitations to our conclusions due to the design of our study. First, because our benchmark data also comes from a survey, we cannot be certain that the bias we report is reflective of the bias from actual appliance ownership in the U.S. We believe that RECS is the best benchmark available for national appliance ownership, but large government probability surveys and administrative records are not always completely accurate (Baker et al. 2013). Second, we cannot be certain that the bias we report for appliance ownership is the same as the bias for other questions in the AMT surveys. Yeager et al. (2011) showed that the accuracy of internet surveys can vary across questions within a single survey. Third, the bias reported in this study may not reflect the bias from other AMT surveys or surveys performed with a different online survey method, because the accuracy of a single measure can vary between online surveys (Yeager et al., 2011). Fourth, the bias we report for questions on appliance ownership may be different than the bias for other types of questions, such as personal opinions or questions requiring more technical knowledge.

Another limitation to consider is that RECS was intended to measure households in 2009, while our AMT surveys were performed in 2012. This leads to a question of whether the product ownership observed in RECS 2009 is appropriate to benchmark 2012 survey data. We used RECS 2009 as the benchmark for our surveys because it is the most recent

survey representing the appliance usage of all U.S. households. To determine whether ownership was likely to change between 2009 and 2012, we examined the trends in product ownership from past RECS surveys (from 1993 to 2009 with a four-year interval). Refrigerators and freezers have had fairly consistent penetration rates since 1993, with less than a one percent change in all categories of ownership for these products between 2005 and 2009. Therefore, it is unlikely that ownership would have changed drastically between 2009 and 2012. The average number of ceiling fans owned by households with at least one ceiling fan increased from 2.5 in 1997 to 2.9 in 2009, but ownership of ceiling fan remained flat between 2005 and 2009. Hence, it is also unlikely that there was a sudden change in ceiling fan ownership between 2009 and 2012.

Compared to the relatively static technology seen in refrigeration products and ceiling fans, televisions exemplify the ever-growing demand of more innovative products in the consumer electronics market. The television industry has gone through a phase-out of cathode ray tube (CRT) technology and the introduction of flat panel displays during the past few years. The average number of televisions per households with at least one television has increased from 1.9 in 1993 to 2.6 in 2009. During the same time, the fraction of households owning one television dropped by half and the fraction of households owning more than two televisions grew steadily. Even though more households own more televisions in the late 2000s as compared to the early 1990s, the trend seemed to reach a plateau recently: the average number of televisions per household increased by less than 0.05 between 2005 and 2009. Nonetheless, the fast-changing nature of television innovation and consumer preference may be one of the reasons we found televisions to have the greatest absolute difference between AMT and RECS after weighting. Based on these observations, we believe that RECS 2009 is a reasonable benchmark to use for our AMT surveys. At the same time, we cannot be certain that the bias we report is the same as the bias we would find if we used to a benchmark survey performed in 2012.

What value are the results of this study for researchers and policy-makers working in the area of energy efficiency? By showing that online surveys of appliance usage can accurately estimate appliance ownership, the present study suggests that online surveys can be used to estimate a host of important aspects about appliance usage that would otherwise be unknown when doing energy efficiency research. For example, consider a recent study on ceiling fans (Kantner et al. 2013). Ceiling fans are a common appliance in households, but there has been little study of their energy use. Through an online survey, the study's authors were able to estimate the frequency with which ceiling fans were operated throughout the year across the U.S., the distribution of speeds that the fans were operated at, and the proportion of these fans that have lights. Knowledge of this information allows a much more precise view of ceiling fan and ceiling fan lit kit energy use than was possible with pre-existing data sources. As a result, researchers can now perform better cost-benefit analyses of energy efficiency policies for these appliances. Another example is a study that focused on refrigeration products other than traditional refrigerators and freezers, such as wine chillers and residential ice-makers (Greenblatt et al. 2013b). In this case, there was even less information available than for ceiling fans—RECS does not even track these products. Using an online survey, the study authors

were able to estimate the saturation, size, lifetime, and refrigeration technology of many of these products. Each of these parameters are important for estimating (a) the current energy consumption of these products, and (b) the potential impact of energy efficiency policies focused on these products. Prior to that study, none of those parameters were available, so there was no data upon which researchers could provide analysis to guide policy decisions.

Based on the results of the current study, what is the appropriate role for online surveys of appliance usage? A recent task force report on non-probability surveys published by the American Association for Public Opinion Research highlighted the important trade-off between accuracy, timeliness, and cost in sample surveys (Baker et al. 2013). Government statistical agencies often need to focus on accuracy, which necessitates the use of large probability surveys. Other researchers, however, are willing to sacrifice some accuracy in order to get robust results within time and budget constraints. Many important appliance studies have either relied on large surveys (e.g., the California Residential Appliance Saturation Survey[15] and Commercial End Use Survey[16]; the Commercial Buildings Energy Consumption Survey[17]; Parekh et al. 2012; and Urban et al. 2011), extensive field monitoring of equipment (e.g., Lanzisera et al. 2013; Greenblatt et al. 2013a; and Mercier & Moorefield 2011), or some combination of both (e.g., the Residential Building Stock Assessment[18]; Zimmermann et al. 2012; and Bensch et al. 2010). All of these studies have provided very valuable data, but at considerable expense and requiring a significant amount of time to execute. The AMT data presented in this paper highlight a complementary approach for gathering high-quality appliance information very quickly at significantly reduced cost. Although this approach is not suitable for every application, we believe that online surveys with proper demographic weighting can complement existing data sources, fill gaps in data, and help guide preliminary policy decisions.

5. Acknowledgments

Bereket Beraki, Sarah K. Price, Stacy Pratt and Henry Willem provided an invaluable contribution to this project by managing the collection of Amazon Mechanical Turk data. Andrea Alstone, Mia Forbes Pirie, Mohan Ganeshalingam, Karina Garbesi, Samantha Infeld, Colleen Kantner, Erik Page, Alex Valenti, and Vagelis Vossos provided assistance in developing and executing the surveys. Gregory Rosenquist and Alex Lekov provided high-level support and encouragement. We thank the U.S. Department of Energy, Building Technologies Office for financial support.

[15] http://www.energy.ca.gov/appliances/rass/
[16] http://www.energy.ca.gov/ceus/
[17] http://www.eia.gov/consumption/commercial/index.cfm
[18] http://neea.org/resource-center/regional-data-resources/residential-building-stock-assessment

6. References

Attari, S.Z., DeKay, M. L., Davidson, C. I., Bruine de Bruin, W. (2010). Public perceptions of energy consumption and savings. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 107(34), 16054-16059.

Attari, S.Z. (2013). Perceptions of water use. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 111(14), 5129-5134.

Bailey, B. J. R. (1980). Large sample cell weighting confidence intervals for the multinomial probabilities based on transformation of the cell frequencies. *Technometrics*, 22, 583-589.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., … Zahs, D. (2010). AAPOR Report on Online Panels. *Public Opinion Quarterly*, 74(4), 711–781.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., Tourangeau, R. (2013). Report of the AAPOR task force on non-probability sampling. American Association for Public Opinion Research, Deerfield, IL, USA.

Baker, R. and Le Guin, T. D. (2007). Separating the wheat from the chaff: Ensuring data quality in internet samples. In Ed. Trotman, M. The challenges of a changing world. Proceedings of the fifth ASC international conference, Southhampton, U.K.

Battaglia, M.P., Izrael, D., Hoaglin, D. C., and Frankel, M. R. (2009). Practical considerations in raking survey data. *Survey Practice,* 2(9).

Bensch, I., Pigg, S., Koski, K., and Belshe, R. (2010). Electricity Savings Opportunities for Home Electronics and other Plug-In Devices in Minnesota Homes: A technical and behavioural field assessment. Energy Center of Wisconsin.

Bethell, C., Fiorillo, J., Lansky, D., Hendryx, M., & Knickman, J. (2004). Online consumer surveys as a methodology for assessing the quality of the United States health care system. *Journal of Medical Internet Research*, 6(1), e2.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1): 3–5.

Cherry, S. (1996). A Comparison of Confidence Interval Methods for Habitat Use-Availability Studies. *Journal of Wildlife Management,* 60(3): 653-658.

Cobanoglu, C., Warde, B., Moreo, P. J. (2001) A Comparison of Mail, Fax, and Web-Based Survey Methods. *International Journal of Market Research*, 43(4)

Couper, M. (2000). Web Surveys: a review of issues and approaches. *Public Opinion Quarterly,* 64:464-494.

Deville, J.C., Särndal, C.E., and Sautory, O. (1993) Generalized raking procedures in survey sampling. *Journal of the American Statistical Association,* 88(423): 1013-1020.

Goodman, J., Cryder, C., Cheema, A. (2013) Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral. Decision. Making*, 26: 213–224.

Gosling, S. D., Vazire, S., Srivastava, S., John, O. P. (2004) Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions about Internet Questionnaires. *American Psychologist*, 59(2): 93-104

Greenblatt, J.B., Pratt, S., Willem, H., Claybaugh, E., Desroches, L.-B., Beraki, B., Nagaraju, M., Price, S.K., and Young, S.J. (2013a). Field data collection of miscellaneous electrical loads in Northern California: Initial results. Lawrence Berkeley National Laboratory, LBNL report number 6115E, February. http://escholarship.org/uc/item/5cq425kt. Accessed 20 April 2014

Greenblatt, J.B., Young, S.J., Yang, H.C., Long, T., Beraki, B., Price, S.K., Pratt, S., Willem, H., Desroaches, L.B., and Donovan, S.M. (2013b). U.S. residential miscellaneous refrigeration products: results from Amazon Mechanical Turk surveys. Lawrence Berkeley National Laboratory, LBNL report number 6537E, November.

Greenblatt, J.B., Yang, H.C., Desroaches, L.B., Young, S.J., Beraki, B., Price, S.K., Pratt, S., and Willem, H. (2013c). U.S. residential consumer product information: Validation of methods for post-stratification weighting of Amazon Mechanical Turk. Lawrence Berkeley National Laboratory, LBNL report number 6163E, March.

Hicks, A. L., Theis, T. L. (2013). Residential energy-efficient lighting adoption survey. *Energy Efficiency*, DOI: 10.1007/s12053-013-9226-6

Ipeirotis, P. (2010). Demographics of Mechanical Turk. http://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html. Accessed 20 April 2014

Kalton, G., and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics,* 19(2): 81-97.

Kantner, C. L. S., S. J. Young, S. M. Donovan, K. Garbesi, S. Pratt, H. Willem, B. Beraki, and S. K. Price (2013). Ceiling Fan and Ceiling Fan Light Kit use in the U.S.—Results of a Survey on Amazon Mechanical Turk. Lawrence Berkeley National Laboratory, LBNL Report number 6332E, July. http://escholarship.org/uc/item/3r67c1f9. Accessed 2 October 2013.

Lanzisera, S., Dawson-Haggerty, S., Cheung, H.Y.I., Taneja, J., Culler, D., and Brown, R. (2013). Methods for detailed energy data collection of miscellaneous and electronic loads in a commercial office building. *Building and Environment,* 65, 170-177.

Loosveldt, G. and Sonck, N. (2008). An Evaluation of the Weighting Procedures for an Online Access Panel Survey. *Survey Research Methods*, 2, 93-105.

McNary, B., and Berry, C. (2012). How Americans are Using Energy in Homes Today. ACEEE Summer Study 2012, 12-17 August, Pacific Grove, CA. http://www.aceee.org/files/proceedings/2012/data/papers/0193-000024.pdf. Accessed 20 April 2014.

Mercier, C., and Moorefield, L. (2011). Commercial Office Plug Load Savings and Assessment. California Energy Commission, PIER Energy-Related Environmental Research Program. CEC-500-08-049.

Nielsen (2012). Nielsen Television Audience Measurement Data, 2009-2012, The Nielsen Company, LLC.

Paolacci, G., J. Chandler and P. G. Ipeirotis (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5 (5): 411-419.

Parekh, A., Wang, P., and Strack, T. (2012). Survey Results of User-Dependent Electricity Loads in Canadian Homes. 2012 ACEEE Summer Study on Energy Efficiency in Buildings.

Steel, R. G., Torrie, J. H., Dickey, D. A. (1996). *Principles and Procedures of Statistics: A Biometrical Approach.* McGraw-Hill.

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys* (p. 198). New York: Oxford University Press.

Urban, K., Tiefenbeck, V., and Roth, K. (2011). Energy consumption of consumer electronics in U.S. homes in 2010. Fraunhofer Center for Sustainable Energy Systems.

U.S. Department of Energy: Energy Information Administration (2011). How does EIA estimate energy consumption and end uses in U.S. Homes. http://www.eia.gov/consumption/residential/reports/2009/methodology-end-use.cfm. Accessed 12 March 2014.

U.S. Department of Energy: Energy Information Administration (2012). Residential Energy Consumption Survey: 2009 RECS Survey Data. http://www.eia.gov/consumption/residential/data/2009/. Accessed 20 April 2014.

U.S. Department of Energy: Energy Information Administration (2013a). Annual Energy Outlook: Residential sector key indicators and consumption. http://www.eia.gov/oiaf/aeo/tablebrowser/#release=AEO2011&subject=0-AEO2011&table=4-AEO2011&region=0-0&cases=ref2011-d120810c. Accessed 24 February 2014.

U.S. Department of Energy: Energy Information Administration (2013b). Residential Energy Consumption Survey (RECS) 2009 Technical Documentation Summary. U.S. Department of Energy, Washington D.C.

Van Ryzin, G. G. (2008). Validity of an On-Line Panel Approach to Citizen Surveys. *Public Performance & Management Review*, 32(2), 236–262.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levindusky, M. S., Simpser, A., and Wang, R. (2011). Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Constructed with Probability and Non-Probability Samples. *Public Opinion Quarterly*, 75(4): 709-747.

Zeifman, M., Roth, K. (2011). Nonintrusive appliance load monitoring: Review and outlook. *IEEE Transactions on Consumer Electronics*, 57(1): 76-84.

Zimmermann, J.-P., Evans, M., Griggs, J., King, N., Harding, L., Roberts, P., and Evans, C. (2012). Household Electricity Survey: A study of domestic electrical product usage. Intertek Report R66141, May.

7. Appendices

7.1 Difference in Phrasing Demographic Questions between AMT and RECS

Firstly, the gender question asked for the head of household in RECS, while the AMT survey asked for the gender of the participant. Secondly, the response categories for household income were grouped differently in the set-top box survey, compared to ceiling fans and refrigeration products, and all three of these were grouped differently from RECS. Thirdly, in the AMT survey there are a set of age questions "How many household members are aged 0-19?", "How many household member are aged 20-29?" and so on, with the final question being "How many household members are aged 70 or more?" The response options ranged from "1" to "10 or more". In RECS the question is phrased "What age is the householder?", "What age is the second household member?" and so on up to 15 possible household members. The response categories are more refined, grouping ages by five years, rather than 10 years as in the AMT survey and the final category is "85 or more."

7.2 Determining Order of Demographic Variables for Cell Weighting

The cell weighting method was developed by directly comparing groups of demographic variables to RECS. The method began with a single demographic group for example census region. Each response in our survey was assigned one of ten possible weights, depending on which census region the participant was from. The next step added a second demographic variable while retaining the first; e.g., for instance, at the first step, all responses from New England are assigned weight A. At the second stage all responses from New England will be assigned weight A×F or A×M, depending on whether the respondent is Female (F) or Male (M). This process was then repeated by adding more demographic variables, until a final combined weight was obtained for each demographic combination.

In determining the appropriate number of demographic variables to be included in the cell-weighting method, Table 3 shows the percentage of responses with no match in RECS for various combinations of demographics. The method was tested with up to six variables (region, people per household, race, education, number of people aged 20-29 years and income). However, it was found that at this point, approximately 50% of the survey responses had no corresponding RECS value, so the value of adding a sixth variable was marginal. Therefore, a maximum of five variables was used: region, people per household, race, education and number of people aged 20-29 years (hereafter "N20-29").

Table 3: Percentage of AMT survey responses with no corresponding response in RECS

Region	No. of occupants	Race	N20-29	Education	Income	TS Survey	CF Survey	RP Survey
1						0.00%	0.00%	-
1	2					0.57%	0.00%	-
1	2	3				1.88%	2.02%	-
1	2	3	4			3.81%	4.57%	-
1	2	3		4		5.83%	6.27%	-
1	2	3	4	5		16.08%	17.88%	-
	1					0.48%	0.00%	0.10%
2	1					0.83%	1.13%	0.56%
2	1	3				2.37%	3.29%	2.38%
2	1	3	4			6.00%	7.22%	5.73%
2	1	3	4	5		-	22.34%	19.23%
2	1	3	4		5	-	29.27%	23.24%
2	1	3	4	5	6	-	53.95%	48.30%

Notes: Not all combinations were applied to all data sets. Those that were omitted are represented by "-".The number indicates the order in which the cell-weighting method was applied to each demographic variable. Survey codes: TS = televisions and set-top boxes; CF = ceiling fans; RP = refrigeration products.

The analysis of null responses allowed us to identify which demographic variables to use as well as the order of the variables. Region and people per household were generally the

most answered with a mean of only 0.2% and 0.1% null responses respectively; therefore these variables were chosen as the first two. Among the rest of the demographics, race (combined with Hispanic) and education had slightly higher numbers of null responses, at 1.4% and 0.8% respectively. Income had the highest null response rate, with an average of 5.5% across all the surveys; as a result, we elected not to use this variable in the cell weighting method. It was also found that income is highly correlated with education; therefore it was not considered necessary to use both. Although the gender question received fairly high response rate, it was discounted as it had the most similar distribution to RECS when unweighted and added no additional value by including it. For ages, it was possible for a household to have members from more than one category, therefore to compare to RECS we looked at households that had at least one member from a given age category. In doing this we found highly significant over-representation of 20-29 year olds, and therefore decided to incorporate this into the cell weighting method, by dividing the responses to the 20-29 year old question into four categories (none, 1, 2, 3 or more). The order of the subsequent variables used in cell weighting was then determined to be race and N20-29 through trial and error. The results of this were found to be similar for all three products; therefore only one example (television and set-top box survey) is used in the following discussion.

Fig. 3 shows the difference between the weighted AMT results and RECS for the television and set-top box survey, using two different orderings of demographic variables. Both orderings began with region (1st) and number of occupants (2nd), but the first ordering followed with race (3rd) and education (4th), shown in the Figures in black; while the second ordering had education (3rd) and race (4th), shown in gray. Not surprisingly, using race before education resulted in better agreement with RECS in the race categories, while using education before race resulted in better agreement in the education categories. However, the scale of improvement of the race categories when race was first was much more significant than when education was first. Therefore, using race first gave a better fit overall. It is also interesting to note that using race first improved agreement for the income categories to a greater extent than using education first.

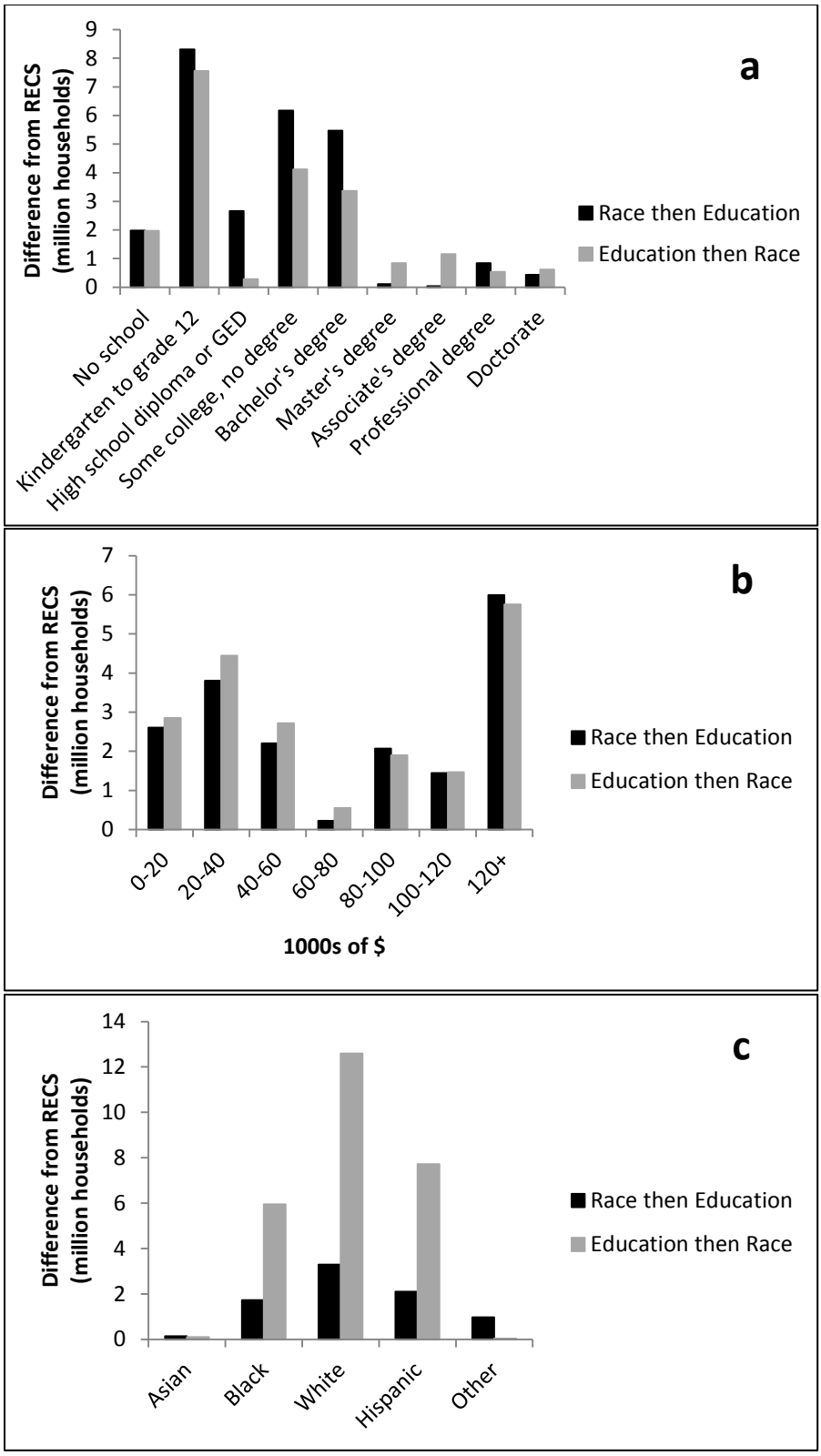
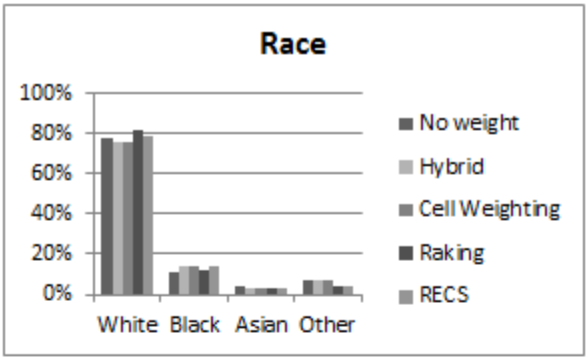
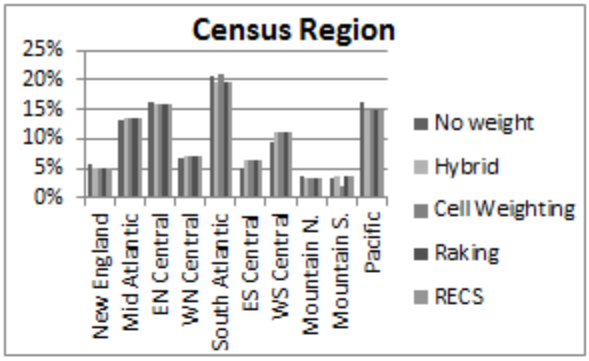
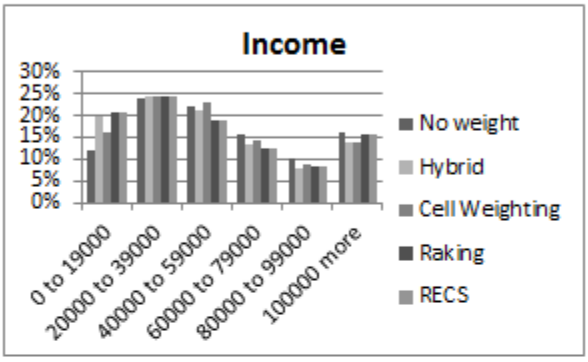


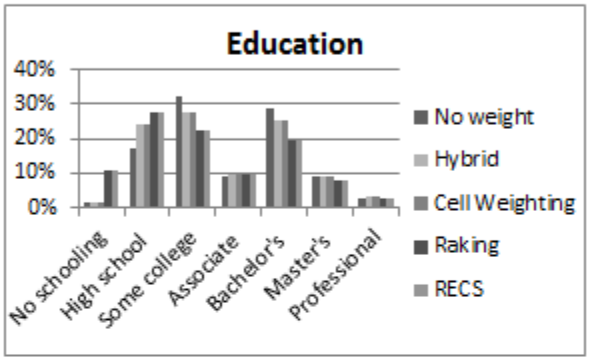
Fig. 3 Comparison of difference from RECS for the a) education categories; b) income categories; c) race using different orders of demographics (television and set-top box survey example). For context, the assumed number of U.S. households was 116 million.

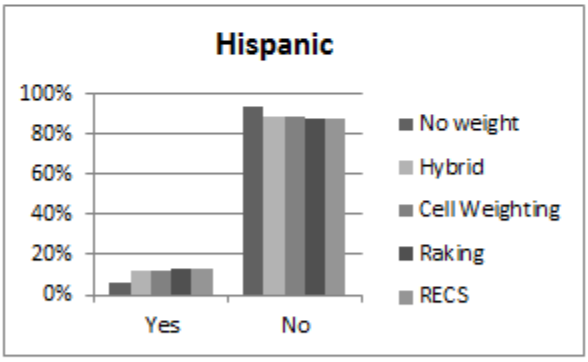
7.3 Demographic Distribution Comparison for Refrigeration Products

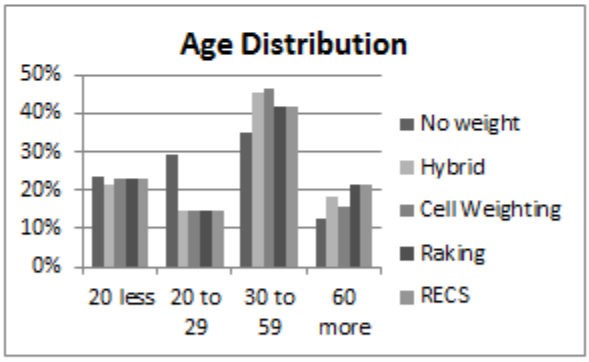


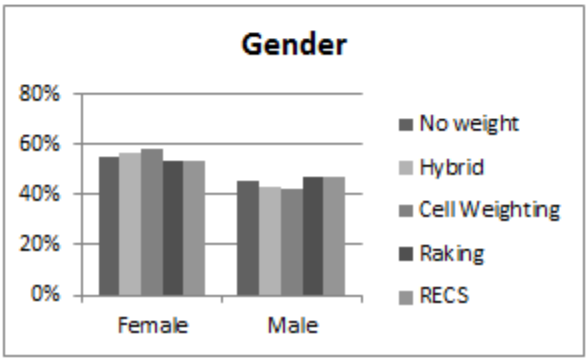


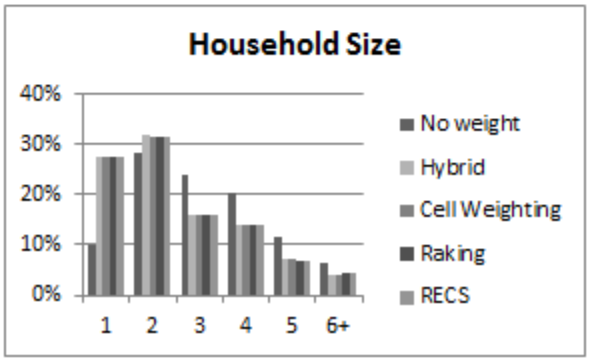












7.4 Product Ownership Questions

Survey Product	RECS	Amazon Mechanical Turk
Refrigerators	How many refrigerators are plug-in and turned on in your home?	How many refrigerators are plugged in at your home right now?
Freezers	How many separate freezers are used in your home?	How many stand-alone freezers are plugged in at your home right now?
Televisions	How many televisions are plugged-in in your home?	How many TVs (in working order) does your household own?
Ceiling Fans	How many ceiling fans does your household use?	How many ceiling fans (both indoor and outdoor) are there in your home?