Smart Prosperity Institute

CLEAN ECONOMY WORKING PAPER SERIES

October 2020 / WP 20-10

A systematic review of energy efficiency home retrofit evaluation studies

Lauren Giandomenico
University of Ottawa

Maya Papineau
Carleton University

Nicholas Rivers
University of Ottawa

The opinions and findings herein are the author's alone, and do not necessarily represent those of their academic institution or Smart Prosperity Institute.

This research project was supported by Smart Prosperity Institute's Economics and Environmental Policy Research Network (EEPRN) and the Greening Growth Partnership

Ce projet a été réalisé avec l'appui financier de :
This project was undertaken with the financial support of:

Environnement et Changement climatique Canada
Environment and Climate Change Canada

SSHRC CRSH

uOttawa
Institut de l'environnement
Institute of the Environment

The Clean Economy Working Paper Series disseminates findings of ongoing work conducted by environmental economic and policy researchers. These working papers are meant to make results of relevant scholarly work available in a preliminary form, as a way to generate discussion of research findings and provide suggestions for revision before formal publication in peer-reviewed journals.

A systematic review of energy efficiency home retrofit evaluation studies[1]

Lauren Giandomenico, University of Ottawa
Maya Papineau, Carleton University
Nicholas Rivers, University of Ottawa

This draft: October 2020

Abstract

This is the first systematic review of studies that evaluate the energy savings and cost effectiveness of residential energy efficiency retrofit programs. We specifically review 33 evaluations of 19 residential retrofit programs that were implemented in the United States and Europe between 1979 and 2014. Our sample is restricted to program evaluations that used actual household billing data from 159,935 retrofitted households. We report four main findings. First, none of the studies in our sample reported deep savings (e.g., 50% or greater) from retrofit programs. The mean reduction in measured electricity and/or fuel consumption due to energy efficiency retrofits for all programs included in our sample was roughly 7.5%. However, because many households use fuel and electricity, total household energy savings from retrofit programs evaluated in our sample are probably smaller. Second, reported program savings decreased as the internal validity of study design increased. Third, in terms of realized savings and cost-effectiveness, the most promising retrofits were water heater insulation and programmable thermostats, and the least promising retrofits were storm windows and doors. Fourth, programs with high reported savings and low costs of conserved energy served low-income, fuel-heated households exclusively.

1.0 Introduction

The global residential building stock accounts for 28% of the world's energy consumption, and is a focal point for imminent climate policy action (International Energy Agency, 2019). Recent projections from the U.S Energy Information Administration (EIA) anticipate an increase in world energy use by 50% by 2050, 65% of which they attribute to worldwide residential and commercial building energy consumption (EIA, 2019). Given the long lifespan of buildings, most of the residential buildings responsible for this anticipated increase in emissions have already been built; by 2050, the United Nations projects that 65% of the current residential building stock will still be in use (United Nations Environment, 2017). Any long-term climate mitigation strategy, and especially those which commit to net-zero-emissions by 2050 (United Nations Climate Change, 2019), therefore needs to include measures to reduce greenhouse gas (GHG) emissions from the existing residential building stock.

[1] We gratefully acknowledge funding from the Smart Prosperity Institute and the Canada Research Chairs program at the Social Sciences and Humanities Research Council.

Residential energy efficiency retrofit programs are widely employed means to achieve energy savings in existing housing for four main reasons. First, residential retrofit programs are seen as a cost-effective approach to GHG mitigation. By targeting the residential buildings that are proverbial “low-hanging-fruit”-- inefficient homes that require only minor, low-cost retrofits to improve energy conservation-- it is argued that retrofit programs can reduce GHG emissions at low cost (Burgett & Chini, 2013). Second, a collection of behavioural and market failures[2] that can deter households from retrofitting their homes on their own may justify government intervention (Gillingham and Palmer, 2014). Third, residential energy efficiency retrofit programs are favoured during times of economic hardship because they stimulate the economy by creating jobs in the construction sector (Bell, 2012). Following the 2008/2009 recession, green stimulus packages that included building retrofits were implemented in China, Korea, Japan, the European Union (including many of its member states), and the United States (IEA, 2020). Most notably, the United States American Recovery and Reinvestment Act (ARRA) increased funding to a national low-income retrofit program, the Weatherization Assistance Program (WAP), by USD 5 billion; France and Germany received EUR 400 million and EUR 3.3 billion from the European Energy Efficiency Fund, respectively, to improve their statewide retrofit programs; and Korea allocated USD 6 billion (80% of their total green stimulus package) to improve the energy efficiency of buildings (IEA, 2020). Fourth, residential retrofit programs are popular policy choices. Governments around the world have framed residential retrofit programs as mechanisms that mitigate carbon emissions and fuel poverty, improve health benefits, create jobs, and/or support economic growth (Kerr et al., 2017). These interests connect a wide range of political actors, lobbyists, and interest groups, which further motivates residential retrofit programs as a suitable choice for policy action (Kerr et al., 2017).

Residential energy efficiency retrofit programs typically work by offering partially or fully-subsidized energy efficiency upgrades to homes in need of repairs. Eligible households are usually required to participate in a home-energy audit to assess the building’s efficiency, and are then recommended a list of energy efficient upgrades based on their energy efficiency rating. Examples of building retrofits commonly incentivized under a residential retrofit program include attic or wall insulation, the installation of storm windows and doors, upgrades to heating or cooling systems, and caulking or weatherstripping.

Despite their popularity, mixed results from residential retrofit program evaluations underscore that not all retrofit programs reach the same level of savings (Fowlie et al., 2015); there are factors underlying residential retrofit programs that influence program effectiveness. Given the considerable energy savings potential in the residential sector, a primary motivation for research in this field is to understand how retrofit program design can be improved in order to maximize energy savings while maintaining cost-effectiveness.

Previous studies have made an effort to isolate program attributes that generate large savings by synthesizing results from program evaluation studies, though most of the current literature is

[2] These are often coined the “energy efficiency gap” and include imperfect information, principal-agent dilemmas, credit constraints, regulatory failures, learning-by-using problems, and non-standard beliefs, preferences, and decision-making. See Gillingham and Palmer (2014) for detailed descriptions.

outdated (Stern et al., 1986). Whereas current evaluations rely on metered consumption data in ex post analyses, many earlier evaluations (pre 2000) used simulated household billing data to forecast program savings in ex ante analyses. While some evaluations from this period used real household billing data to estimate program savings, these studies were often smaller, shorter-term, and produced potentially biased estimates of program savings because of selection effects left unaccounted for (Brown & Macey, 1985). Papers that synthesized the literature about program savings from this period are subject to the same flaws (Berry & Johnson 1983; Wall et al., 1983).

Improvements in recent program evaluations have strengthened the credibility of savings estimates reported. Today, program savings are empirically derived - often experimentally or quasi-experimentally - using households' billing data. Recent papers have discussed findings reported in newer evaluation studies, though efforts have not yet been made to synthesize these results systematically (Cluett and Amann, 2016; Gillingham et al., 2018).

This is the first systematic review of residential retrofit evaluation studies, which advances literature about residential retrofit programs in four key ways. First, we compile an exhaustive survey of all primary studies in the peer reviewed literature that evaluate energy efficiency retrofit programs. Second, we systematically review these studies for measures of program savings and cost-effectiveness. Third, we consider how savings estimates and cost-effectiveness change based on program attributes, household characteristics, and study design, in an effort to isolate factors that affect program outcomes. Fourth, we draw lessons from program evaluations and provide recommendations for retrofit program designs that can maximize savings while maintaining cost-effectiveness.

In this study, we specifically reviewed how savings estimates and cost-effectiveness varied by program attributes, study design, and household characteristics, as measured in 33 residential energy efficiency retrofit program evaluation studies in the United States and Europe. Following a detailed literature search, we systematically included peer-reviewed papers that met the following criteria: The study pursued applied research about residential energy efficiency retrofit programs in Canada, Europe, or the US using metered energy consumption data; evaluated program-induced reductions to household electricity and/or fuel consumption; and targeted residential buildings in Canada, Europe or the US. We screened the sample by applying our inclusion criteria to the papers' abstracts, followed by the papers themselves. We then systematically extracted program evaluation outcomes from our study sample. We extracted information according to three primary outcomes of interest: 1) reductions in household electricity or fuel consumption (savings); 2) cost-effectiveness, and 3) findings idiosyncratic to the studies. From this, we determined the mean savings for the study sample, by program, study design, primary heating source, and income characterization. We also discussed the programs' cost-effectiveness, and identified commonalities between programs that we found correlated with higher measured savings and lower costs of conserved energy and payback periods.

This paper has four main findings**.** First, the residential retrofit programs evaluated by studies included in our review delivered modest savings. The mean reduction in measured electricity and/or fuel consumption due to energy efficiency retrofits for all programs included in our sample was about 7.5%. Because many households use fuel and electricity, total household energy savings due to retrofit programs evaluated in our sample are probably smaller than 7.5%. None of the studies in our sample reported deep savings (e.g., 50% or greater) from retrofit programs; the highest program savings reported in any individual study were 26%.

Second, the magnitude of savings reported by studies in our sample depended on the type of study design used to evaluate the residential retrofit programs. Reported program savings decreased as the internal validity of study design increased; savings were roughly 13%, 12%, 11%, and 6.8% for simple difference, cross-section, difference-in-difference, and RCT designs, respectively. This finding reinforces the importance of using study designs with high internal validity to evaluate program savings. The large discrepancy between reported savings using simple difference, cross-section, and difference-and-difference designs compared to the savings reported from RCT designs raises questions about the accuracy of previously-reported savings estimates.

Third, across a variety of methods, water heater insulation and programmable thermostats were found to be the most promising retrofits in terms of savings and cost-effectiveness (Brown & Berry, 1995; Liang et al., 2018). In contrast, retrofits that frequently reported low measures of cost-effectiveness were storm windows and doors (Brown and Macey, 1985; Hirst, 1987). In general, for studies that estimated cost-effectiveness in terms of the cost of conserved energy for energy efficiency retrofits, estimates ranged from US 3 cents per kWh to US 48 cents per kWh.

Fourth, the most successful programs (with high reported savings and low costs of conserved energy) served low-income, fuel-heated households exclusively. Retrofit programs targeted at low-income households were found to reduce energy consumption by 12% on average, which was double the estimate for programs that targeted middle and high income households. Studies and/or programs that targeted fuel heated homes saw 2% and 5% higher savings than studies and/or programs that targeted electrically-heated or electrically and fuel-heated homes, respectively.

This paper proceeds as follows. We first explain the conceptual framework used to guide our review (*s.* 2.0). In Section 3.0, we provide a step-by-step overview of our research method, which includes a definition of the literature search (*s.* 3.1), our inclusion/exclusion criteria (*s.* 3.2), our extraction code used to pull relevant information from the studies included in our final sample (*s*. 3.3), and a synopsis of our systematic extraction and analysis approach (*s*. 3.4). In Section 3.5, we provide an overview of the study designs included in our sample, and discuss the advantages and disadvantages of each. Limitations to our research method are discussed in Section 3.6. We present the results of our analysis in Section 4.0, followed by a discussion section that situates our results in the wider body of literature and highlights areas of future research (*s.* 5.0). Finally, in Section 6.0, we offer concluding remarks.

2.0 Design of the Systematic Review

Following Andor and Fels (2018), we adopted The Campbell Corporation's (2014) PICOS framework to guide our systematic review (See *Table 1*). We defined our population of interest as studies that evaluated residential energy efficiency retrofit programs in Canada, Europe, or the US. We were only interested in studies that evaluated savings in household electricity or fuel consumption as a result of participating in a home retrofit program. Savings are the difference in household electricity and/or fuel consumption after participating in a home retrofit program compared to the level of household electricity and/or fuel consumption that would have occurred in a counterfactual situation in which the household did not participate in the program.

Our secondary outcome of interest were measures of cost-effectiveness, like the cost of conserved energy and payback period. We also considered other indicators of program effectiveness, like the internal rate of return, Net Present Value, the realization rate, and findings idiosyncratic to the evaluation studies.

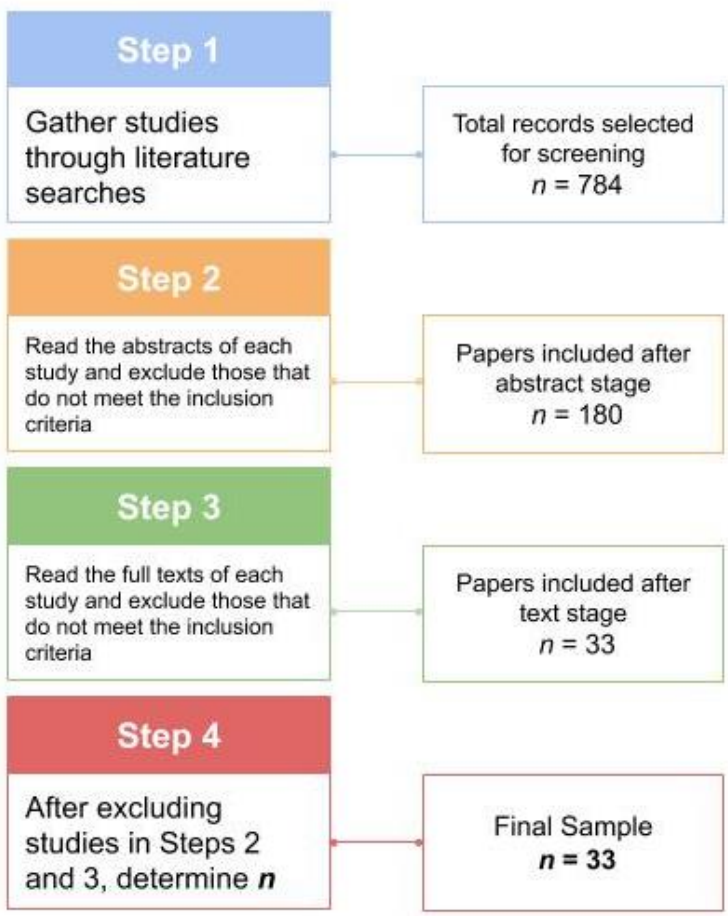
Table 1: Application of the PICOS framework to the current study

PICOS definition	Application of *PICOS* to the current study
Population of interest with a given condition or concern	Residential buildings located in Canada, Europe, or the US
Intervention of interest used to address the condition or concern	Residential energy efficiency retrofit programs
Comparison used to isolate the effect of the intervention on the population and condition of interest	1) Household energy consumption before and after retrofit program participation 2) Household energy consumption in homes that have participated in a retrofit program compared to those that haven't
Outcomes on which improvement should appear	<u>Primary outcomes:</u> household metered energy use (electricity, fuel, or both) <u>Secondary outcomes</u>: measures of cost-effectiveness <u>Tertiary outcomes</u>: Internal rate of return, Net Present Value, realization rate, welfare effects, and other outcomes that are idiosyncratic to the particular study.
Study designs used to evaluate the effects of the intervention on outcomes	RCTs, matching, difference-in-difference, instrumental variable estimation, and regression discontinuity design.
Source: *Campbell Systematic Reviews: Policies and Guidelines. Campbell Systematic Reviews 2014: Supplement 1*	

3.0 Research Method

Figure 1 outlines our sample selection procedure, which we explain in further detail below.

Figure 1: **Procedure for determining the study sample**



3.1 Step 1 - Define the Literature Search

We developed a systematic combination of keywords that we used to search a predetermined set of scholarly databases. The systematic keyword search and the databases upon which the literature search were employed is presented in *Table A* in the Appendix. The search results returned a total of 784 abstracts.

3.2 Step 2 - Develop and Apply Inclusion/Exclusion Criteria to Abstracts

We systematically applied the inclusion criteria checklist to each of the 784 abstracts captured in the database search. Of these, 180 papers were included into the next screening phase of our review. For inclusion, each peer-reviewed study needed to evaluate residential energy efficiency retrofit program using metered energy consumption data, refer to at least one of the outcome variables (see *Table 1*), and target residential buildings in Canada, Europe or the US. A copy of the inclusion checklist is provided in the Appendix (*Table B*). Papers that were included based on this checklist for abstracts were later read through in full and screened again before making it into the final sample of papers. This process is described further in the following section (Step 4). Further details about the selection process are included in the Appendix under Section 5.1: Sample Selection.

3.3 Step 3 - Determine the Final Sample

We determined our final sample by reading the 180 papers that were included from our initial sample collection and by excluding the ones that didn't meet our study criteria. We followed a two-step procedure when reading each of the papers. First, we made sure that the paper itself met our initial inclusion criteria; those that didn't were excluded. Second, we excluded papers that did not report on savings post-retrofit. This additional exclusion criteria was intended to improve the comparability between the studies included in our review.[3] Following this, 33 papers were included in the final sample.[4]

The 33 studies included in our review covered 19 different residential retrofit programs. The studies evaluated programs using data on 159,935 households that received retrofits between 1979 and 2014 in the United States and Europe. No evaluations of Canadian retrofit programs met our inclusion criteria. A list of studies included are provided in the Appendix.

3.4 Step 4 - Develop the Data Extraction Code

Preparing the code sheet was an iterative process; data extraction from the studies and code development occurred simultaneously. We read the longest, most detailed papers first in an attempt to find common themes, measurements, and outcomes that other papers might have also reported on. Because of the unique characteristics of each study, we intended to capture both quantitative (i.e. savings in %) and qualitative information (i.e. the study design) so that the subtleties of each paper could be reflected in the dataset.

Since the retrofit studies reported savings differently, some data cleaning was required. Details about data cleaning are provided in the Appendix in Section 5.2. Overall, we computed weighted averages to obtain mean savings for the entire sample, per program, by income, study design, and by primary heating source.

3.5 Discussion about Retrofit Evaluation Study Designs

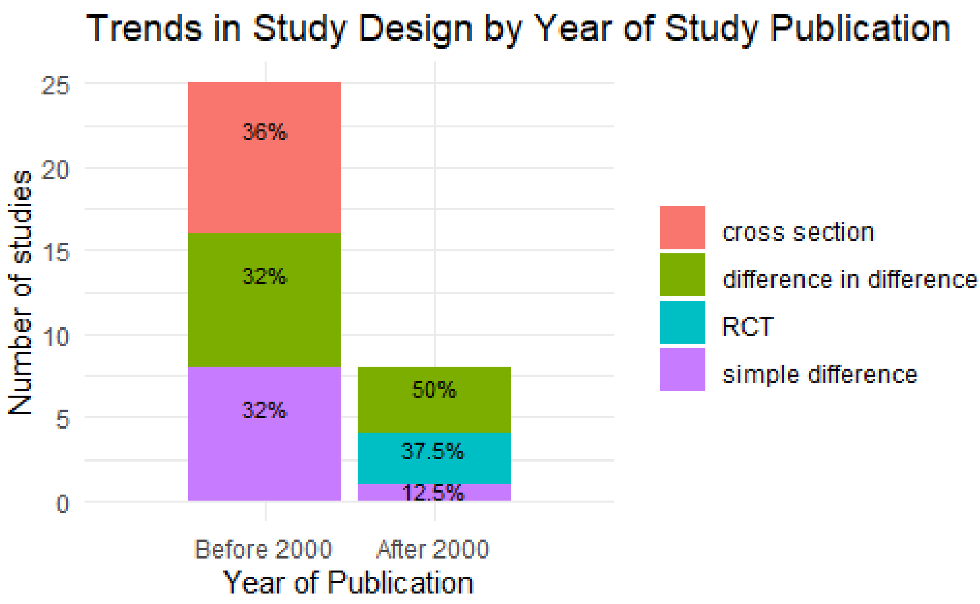
The internal validity of the study design and statistical methods used in retrofit evaluation studies have evolved tremendously over the past several decades. Our study sample reflects this evolution *(Figure 2*), and we make an effort in this study to explore how the evolution in these changes in study design affect the estimates of programmatic savings.

---

[3] We found that many studies initially included met at least one of the relevant outcomes (i.e. The study measured the rebound effect), but did not report on our primary outcome of interest: savings.

[4] Sample attrition of this kind has been observed in comparable studies for similar reasons. See Delmas et al. (2013) as an example.

Figure 2- Trends in the study designs used in papers included in our sample over time



Bar labels denote the prevalence (%) of different study designs included in our sample that were published before and after 2000
N = 33 studies

3.5.1 Simple Differences

Beginning in the late 1970s, residential retrofit program evaluations often ascertained savings by comparing the difference in energy consumption of households before and after receiving a home retrofit, otherwise known as a simple-difference technique. These study designs can be implemented at low cost. Most studies that use this design in our sample use a small number of homes in the program evaluation. These studies do not employ a control group to measure changes in energy consumption, and do not account for households' voluntary decisions to retrofit their homes. The heterogeneity between households that choose to retrofit their homes and those that don't raise concerns about the validity of these evaluation estimates.

3.5.2 Cross-Sectional Studies

Other early evaluation studies employed cross-sectional study designs, which compare a group of houses that are treated by a retrofit program to a control group that do not participate in the program to measure savings. In the case of cross-sectional retrofit evaluation studies, control groups are typically a group of eligible non-participants or future program participants. Concerns about heterogeneity between households that self-select into residential retrofit programs and those don't make future program participants a stronger comparison group than non-participants. However, biased savings estimates are still possible, given the absence of randomization and the fact that unobserved variables may determine program participation and timing. Some cross-sectional studies also make an effort to control for selection into treatment through propensity score matching, or by modelling households' selection into retrofit programs using Inverse Mills Ratios (IMRs). The effectiveness of these approaches is limited because only observed confounding variables can be used to model selection in treatment. Cross-sectional designs are also disadvantaged by their inability to compare treatment effects over time.

3.5.3 Difference-in-Difference

Evaluation studies later refined their measures of programmatic savings by using difference-in-difference approaches. Typically, difference-in-difference designs compare the reductions in electricity or fuel consumption over time between residential buildings that did and did not receive energy efficiency retrofits. While this approach reveals the association between retrofit participation and energy consumption over time, savings estimates are still vulnerable to bias. First, program participation may serve to shift the timing of an investment, but not the investment itself (Rivers and Shiell, 2015). Second, households may undertake other renovations alongside the targeted investment, which confounds the effect of home retrofits on energy savings. Third, the control group may not be comparable to the treatment group due to participant characteristics unobservable to researchers, and that cannot be controlled for using panel estimators. Like cross-sectional analyses, control groups in difference-in-difference designs are usually eligible non-participants or future participants.

3.5.4 Randomized Control Trials

Randomized Control Trials (RCTs) are the gold standard for residential retrofit program evaluation studies because they hold all of the advantages of a standard difference-in-difference approach while also accounting for self-selection into retrofit programs by randomly assigning the treatment and control groups prior to program roll-out. Programmatic savings ascertained through RCT-designed evaluation studies hold the most internal validity compared to those obtained from the aforementioned study designs. Currently, only three RCTs are included in our study because longitudinal data and funding for RCTs have only recently become available. Despite this, these three studies had a cumulative treatment group size of over 100,000 households, which is the largest treatment group size of all study designs included in our sample.

3.6 Limitations of Research

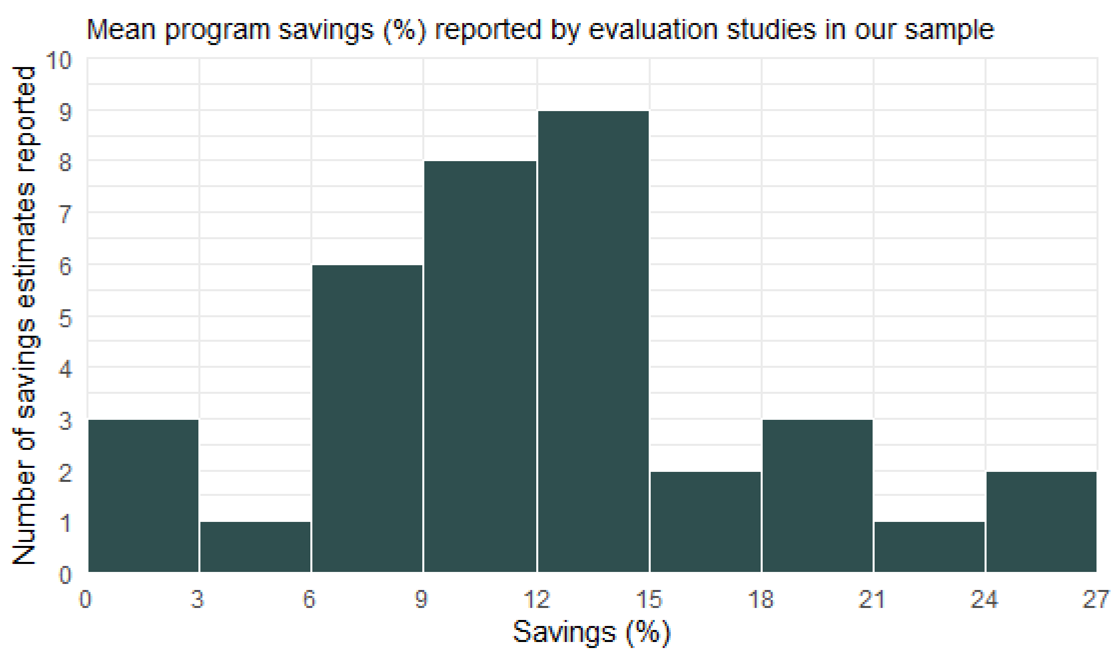
Despite our efforts to understand how program savings estimates vary based on study design, household characteristics, and program attributes, the small number of studies included in our sample limits the external validity of our findings. Program evaluations included our sample may not be fully representative of the outcomes of retrofit programs included in our study, so the findings that we present should be interpreted carefully. There are also a range of residential retrofit programs whose outcomes were not included in our study. We encourage research in this field to continue evaluating retrofit programs so that future systematic reviews can draw stronger conclusions about program outcomes. Finally, since retrofit evaluation studies typically measure the reduction in a specific energy source following retrofit, it is unlikely that the program savings we report represent total household energy savings.

4.0 Results

4.1 Primary Outcome of Interest: Program savings

The average savings for the study sample, and by program and/or study characteristics are captured in *Figure 2* and in *Table 3* located in the Appendix**.** Savings are the difference in household electricity and/or fuel consumption after energy efficiency retrofit program participation compared to the level of household electricity and/or fuel consumption that would have occurred in a counterfactual situation in which the household did not participate in the retrofit program. Because many households consume both fuel and electricity, these energy carrier-specific savings estimates likely overestimate total household energy savings. Studies that did not provide measures of savings in percent (n=4) or baseline energy consumption (n=4) were excluded from the subsample used to calculate mean savings (n=29) as presented in *Figure 2* and *Table 3*. These 29 studies provided 35 measures of program savings, from which mean savings for the study sample were calculated. The weighted average savings for households across all programs measured in our study was 7.5% (weights based on number of retrofitted houses in each study).

Figure 2 - Weighted mean of program savings (%) reported by studies in our sample



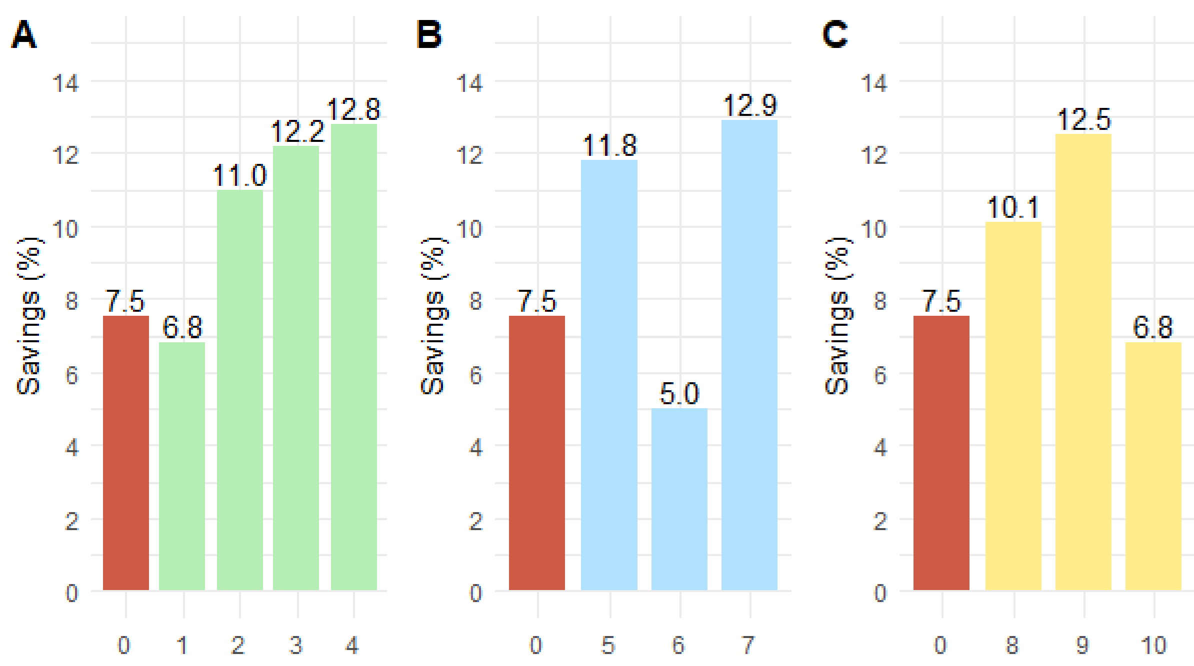
N= 35 measures of savings
Savings are weighted by the size of the treatment group
Savings represent the difference in household fuel and/or electricity consumption between post-retrofit compared to a counterfactual situation in which retrofits were not installed

4.1.1 Variation in Savings by Fuel Type and Income Categorization

Household savings varied based on the characteristics of homes targeted by the study and or program. (*Figure 4* and *Table 4*). Studies and/or retrofit programs targeted fuel-heated homes saw about 2% higher savings than did studies or programs that targeted electrically-heated homes, and 5% higher savings than did studies or programs that targeted both electric and fuel-heated homes (*Table 4*). Moreover, retrofit programs targeted at low-income households saved 7 percentage points more than those targeted at middle-to-high-income households. There were also 13 measures of household savings in which household income was not disclosed; these households observed about 13% in savings. However, it is important to note that the disparities in sizes of the treatment groups and the differences in study designs could affect the internal validity of these findings.

The evaluation study criteria for many studies included in our review restricted their study sample to low-income households living in single-unit dwellings. This potentially excludes a significant population of low-income households from savings evaluations; namely, those living in multi-unit dwellings. Despite the fact that low-income households often live in multi-family-unit buildings, only five program evaluations that measured program outcomes in low-income households included in our study looked at program savings in these buildings. Future studies should investigate further the relationship between tenancy status and program savings.

Figure 4 - Mean savings (%) for the sample by study design, household income categorization, and characteristics of the population of homes that were retrofitted



A - Savings (%) by Study Design	**B- Savings (%) by Household Income** Categorization	**C- Savings (%) by Population of Homes Retrofitted**
0- Sample average 1- RCT 2- Difference-in-Difference 3- Cross-Section	0- Sample average 5- Low income households 6- Middle-to-high income	0- Sample average 8- Electrically-heated 9- Fuel-heated

4- Simple Difference	households 7- Undisclosed income	10- Electrically & Fuel-heated
N = 35 measures of savings Savings are weighted by the size of the treatment group Savings represent the difference in household fuel and/or electricity consumption between post-retrofit compared to a counterfactual situation in which retrofits were not installed		

4.1.2 Variation in Energy Savings by Study Design

It is clear from our study sample that the study design plays a powerful role in the level of savings reported (*Figure 4*).[5] Studies in our sample that measured savings using a simple-difference design found high program savings of 13%, almost double the sample average. Cross-sectional and difference-in-difference designs observed comparable savings, which could reflect the similar types of control groups used in these studies, those which are selected based on observables and are a group of eligible non-participants or future participants. Randomized Control Trial study designs returned the smallest amount of savings at 6.8%. The randomization of treatment and control groups, and the capacity of RCTs to better-control for selection effects could explain the smaller, perhaps more accurate measures of savings.

4.2 Secondary Outcome of Interest: Cost-Effectiveness

Our secondary outcome of interest was the cost-effectiveness of residential retrofit programs. Cost-effectiveness measures for retrofit programs are reported in *Tables 6, 7* and *8*. We interpret each of the columns of these tables below.

4.2.1 Payback Period

The payback periods reported in our sample were calculated using a variety of discount rates and retrofit lifetimes. The resulting payback periods varied widely across programs and by retrofit type, ranging from less than one year to infinity (never paid back).[6] In some studies, the payback period was determined separately for different types of retrofit. The quickest repaying retrofits were low-cost water heater insulation and programmable thermostats at a payback period of less than 4 years (Hirst & Goeltz, 1985). On the contrary, the longest repaying retrofits were: 1) new heating systems, attic/basement/wall insulation, and storm windows at 12-14 years, and 2) storm doors at 18 years (Hirst & Goeltz, 1985).

4.2.2 Cost of Conserved Energy

Seven studies in our sample reported the cost of conserved energy for five retrofit programs (*Table 6*). For comparability purposes, we converted all of the measures into cents per kWh in USD. costs of conserved energy were then adjusted for inflation based on the year at which the costs of conserved energy were calculated. If the year was not provided, we assumed that the cost of conserved energy was calculated for the year of the retrofit.

[5] Further details are provided in *Table 5* in Section 6.3 of the Appendix
[6] Liang et al., (2018) report that at a 5% discount rate, retrofits would never be repaid.

The cost of conserved energy reported in our study sample ranged from 3 to 48 cents per kWh. The mean cost of conserved energy for all measures reported in our study sample (n=7) was about 14 cents per kWh.

4.2.3 Realization Rate

Table 6 also reports the realization rates of the retrofit programs studied. A realization rate of 1 reflects a program in which actual savings equalled predicted savings.[7] A realization rate of <1 represents a situation in which measured savings were less than predicted savings, whereas a realization rate of >1 indicates that actual savings were larger than predicted savings. The realization rates for retrofit programs captured in our study ranged from 25% to 85%; the mean realization rate was 57.92%. This is to say that in every case where a realization rate was computed, engineering estimates of program savings were larger than the savings that were actually observed.

There was a significant discrepancy (45%) between the realization rates for the Colorado and Michigan Weatherization Assistance Programs (WAPs), which is likely a reflection study design. The Colorado study had an attrition rate of 75%, a treatment group of 36 households, did not adjust for self-selection, and did not have a control group to measure savings, which the authors mention could have biased their results (Burch et al., 1993). By contrast, the Michigan study had a treatment group size of 30,000 households, and used quasi-experimental and randomized control designs to measure savings (Fowlie et al., 2015). The large variation in the realization rates for two iterations of the same weatherization program underscores the importance of study design in evaluating program outcomes.

4.2.4 Net Present Value (NPV)

Measures of net present value (NPV) per household, for the region, and for the residential programs overall are presented in *Table 7*. Like the cost of conserved energy and payback periods, various discount rates and projected retrofit lifetimes were used to calculate NPVs. The ranges provided in the table reflect variations in discount rates and retrofit lifetimes. Per household, NPV for all programs with NPV measures captured in our study ranged from -$40 to $4600.

5.0 Tertiary Outcomes of Interest: Trends in Energy Savings by Program and Household Characteristics

After assessing savings and cost-effectiveness, we proceeded to evaluate the studies in our sample for common attributes that resulted in un/favourable program outcomes. We report these findings below.

5.1 The Effect of Baseline Energy Consumption on Savings

[7] Our study sample used engineering estimates as measures of predicted savings.

Three studies (Hirst et. al, 1985; Hirst et al., 1989; Brown & Berry, 1995) specifically identified households' baseline energy consumption as the strongest predictor of savings. Hirst et al. (1985) observed a 0.32 kWh increase in actual saving following a 1 kWh increase in pre-Bonneville Power Administration (BPA) Pilot program electricity use, all else equal. These findings were supported in the BPA Interim programs for the years 1982 and 1983; a 1 kWh/year increase in pre-program energy consumption resulted in an additional 0.3 kWh/year and 0.2 kWh/year in savings for 1982 and 1983 participants, respectively. Likewise, through the Residential Energy Conservation Assistance Program (RECAP), a 1000 kWh increase in pre-retrofit energy consumption saved households an additional 190-200 kWh in savings (Brown & White, 1988). Interestingly, despite the observations of poor savings in mobile homes[8] reported by studies in our sample, once Brown and Berry (1995) adjusted for baseline energy consumption in their model, mobile homes achieved above average gas savings (Goldberg, 1986; Hirst et al., 1989; Tonn et al., 2018)..

The opposite effect, in which homes with better energy efficiency and/or less energy consumption before retrofit had poor savings post-retrofit, further supports these findings (Brown and Macey, 1985). Single family homes that had been retrofitted before participating in the Hood River Conservation Project (HRCP) saved about 1000 kWh less than first-time retrofitters, which speaks to the savings potential of inefficient residential buildings that consume high levels of energy pre-retrofit (Hirst et al., 1989).

However, the perceived positive relationship between baseline energy consumption and household savings could be a reflection of heterogeneity. Goldberg (1986) and Brown and Berry (1995) observed a link between the amount and intensity of retrofit activities applied and households with higher pre-retrofit energy consumption, which insinuates that priority of retrofit resources was given to homes with greater potential for savings, as indicated by pre-weatherized NAC. If inefficient homes were consistently offered more retrofits than efficient ones, the perceived relationship between pre-weatherized NAC and programmatic savings becomes less clear. Moreover, none of the studies that observed a positive relationship between baseline energy consumption and post-retrofit savings randomly assigned treatment and control groups (Hirst et al., 1984; Brown and White, 1988; Goldman and Ritschard, 1986; Brown and Berry, 1995), and only Hirst et al. (1984) controlled for self-selection into the program using an Inverse Mills Ratio. Without random assignment, and considering the few efforts to control for self-selection, it is quite possible that households with inefficient homes were more likely to self-select into the retrofit program than households that did not participate. Consequently, conclusions about the effects of baseline energy consumption on household savings reported in our sample are probably skewed.

5.2 Savings based on the type of retrofit

The effect of different retrofit types on savings was fairly consistent across studies. The installation of storm windows and doors brought about the lowest savings (Hirst et al., 1989; Hirst, 1987; Brown and Macey, 1995), and homes that received insulation measures (Hirst et al., 1989) had the highest savings (Liang et al., 2018; Hirst, 1987; Lee & Englin, 1989; Brown & Berry, 1995). However, Brown and Macey (1985) found that homeowners were less likely to adopt these

[8] There is one exception. One study on the HRCP (Brown et al., 1989) found that low-income households that used electric water heaters to heat their mobile homes had higher savings than other dwellings heated with electric water heaters.

effective insulation measures. They also observed that tax credit incentives were infrequently used to finance insulation measures, and were most often used to fund storm window and door retrofits because of their high cost.

5.0 Discussion

Studies of energy efficiency retrofit programs included in our paper found that electricity or fuel savings by participating homes averaged around 7.5%. Houses that observed the highest savings are those that were low-income, fuel-heated, and had insulation of some kind installed throughout the retrofit process (Brown & Macey, 1985; Liang et al., 2018). For the nine studies that evaluated savings in low-income households, savings averaged 12%, and were 6.5% larger than savings for middle-to-high income households. Programs and/or studies that targeted fuel-heated homes were found to have had 12% in average savings, compared to 10% and 7% savings when programs and/or studies targeted electrically-heated or electrically and fuel-heated homes, respectively. The programs in our sample with the lowest reported cost of conserved energy were those that supplied retrofits to households that met these same criteria.

Our study confirms the observation in a recent systematic review that RCT and quasi-experimental study designs are still far from the norm in residential retrofit program evaluation, despite the recent influx in research in this field (Gillingham et al., 2018). This poses a challenge for the accurate interpretation of savings, especially if studies do not make an effort to control for self selection, ideally through randomization in an RCT. Our paper found a large discrepancy among reported savings when different study designs were used to evaluate residential retrofit programs. Savings decreased as the internal validity of the study design used increased; savings were 13%, 12.5%, 11%, and 6.8% for simple difference, cross-section, difference-in-difference, and RCT designs, respectively. We posit that the RCT's capacity to control for self-selection through randomization played a large role in this difference in savings. Our observation that programmatic savings decreases as more sophisticated controls of self-selection are used challenges Hartman and Doane's (1987) conclusion that the failure to account for self-selection underestimates program-induced savings.

In fact, our review of residential retrofit evaluation studies revealed that some perceived positive relationships between household characteristics and the magnitude of savings reported were likely the result of self-selection effects left unaccounted for. Despite the apparent positive correlation between baseline energy consumption and household savings reported both in the literature (Goldman, 1985; Goldman et al., 1988) and in ten studies included in our sample, none of these ten studies in our review randomized treatment and control groups, which raises questions about the role that households' self-selection into retrofit programs played in this relationship. Since households that voluntarily participate in residential retrofit programs are likely in greater need of retrofits, and are thus more inefficient pre-retrofit than households that choose not to participate, conclusions about household characteristics and program attributes that generate high savings should be viewed with caution, unless self-selection is explicitly accounted for through an RCT.

Only two studies in our sample (both RCTs with treatment groups of 7,500 and 70,000 households) considered how application support and program advertising affected program participation outcomes (Allcott & Greenstone, 2017; Fowlie et al., 2015). Despite in-person application support and, in some cases, expected post-retrofit financial benefits of $1000 per household, both studies observed a participation rate of only 5%. That said, there might be higher savings potential in homes that receive application support than those that are left to apply for retrofit subsidies alone. The 5% of homes that retrofitted following extensive application support in Fowlie et al's (2015) randomized-encouragement evaluation saw about 10% higher savings than homes that retrofitted without support. Similarly, results from Zivin and Novan's (2016) evaluation showed that households who decided to invest in retrofits after receiving behavioural treatments saved 16.9% more energy than households who were not encouraged to participate. Zivin and Novan's (2016) results are about 10% higher than those found in Delmas' et al.'s (2013) meta-analysis about the effect of behavioural interventions on programmatic savings, but nonetheless support the notion that behaviourally-motivated participants might save more energy than those who are left to apply for weatherization on their own. Research should concentrate on how outreach components of retrofit programs can be augmented to maximize participation, and thus, savings.

The striking absence of retrofit program welfare analyses in the studies included in our sample echoes Allcott and Greenstone's (2017) call for a greater emphasis on studying the social costs/benefits of retrofit programs. While the studies we reviewed found low social welfare outcomes (-$2.60 in social welfare per household, and a -9.5% social internal rate of return), the more important takeaway for our purposes is that these findings were derived from only two studies (Allcott & Greenstone, 2017; Fowlie et al., 2015). As countries are likely to ramp up residential retrofit programs in an effort to stimulate the economy amidst COVID-19-induced recessions, it is important that research envision how residential retrofit programs can be adjusted to improve social welfare as well.

6.0 Conclusion

This study systematically reviewed 33 empirical evaluations of savings and cost-effectiveness of residential retrofit programs in the United States and Europe**.** Ultimately, the studies included in our review reported modest fuel or electricity savings. Mean savings for all programs included in our study were 7.5%, none of which reported deep savings (e.g. >50%).

Across a variety of methods, the most promising retrofits appear to be insulation, and the least promising retrofits seem to be the installation of storm windows and doors (Brown & Berry, 1995; Liang et al., 2018; Hirst, 1987). In general, cost-effectiveness in terms of cost of conserved energy reported in seven studies included in our sample ranged from 3 to 48 cents per kWh; the mean cost of conserved energy for the studies was about 14 cents per kWh.

Programs that targeted low income, fuel-heated households tended to have higher savings and also had lower costs of conserved energy. Our systematic review found that studies and/or programs that targeted low-income households saw 12% in savings, which is about twice as much saved than middle and high income households, and about 5% more than the sample average. Moreover, studies and/or programs that targeted fuel heated homes saw 2% and 5% higher savings than studies and/or programs that targeted electrically-heated or electrically and fuel-heated homes, respectively.

We also found that the study design employed in residential retrofit evaluations had a noticeable impact on the magnitude of programmatic savings reported. The gold standard study design for residential retrofit evaluations - the RCT - returned the lowest measure of program savings at 6.8%. This is significantly lower than the 13% savings returned by simple difference designs, and also below the 12% and 11% reported by cross sectional and difference-in-difference designs, respectively. The ability of RCTs to control for heterogeneity by randomizing treatment and control groups is a plausible explanation for these differences in reported savings. Given the discrepancy in savings among study designs, and considering that the RCT is the current gold-standard, it is important that succeeding residential retrofit evaluation studies employ RCT designs to obtain more accurate measures of program-induced savings.

Residential retrofit programs are often reached-for in times of economic hardship because they can reduce GHG emissions in a high-consuming sector while creating jobs and stimulating the economy. Although not all residential retrofit programs are created equal. As presented in this study, some retrofit programs achieved higher savings with better cost-effectiveness than others. An important finding from this review is that a study's measure of program savings is largely dependent on the study design used; not all study designs are created equal, either. Our review underscores that RCTs are the preferred approach to obtain accurate measures of programmatic savings. Research in this field should continue to evaluate the social welfare of retrofit programs and seek to determine an optimal method for targeting and engaging program participants in order to ensure that the maximum number of eligible participants are reached and recruited.

References

Allcott, Hunt, and Michael Greenstone. 2017. "Measuring the Welfare Effects of Residential Energy Efficiency Programs." Working Paper 23386. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w23386.

Andor, Mark A., and Katja M. Fels. "Behavioral Economics and Energy Conservation – A Systematic Review of Non-Price Interventions and Their Causal Effects." *Ecological Economics* 148 (June 1, 2018): 178–210. https://doi.org/10.1016/j.ecolecon.2018.01.018.

Bell, Casey J. 2012. "Energy Efficiency Job Creation: Real World Experiences," American Council for an Energy-Efficient Economy. 21.

Brown, M A, and D L White. 1988. "Stimulating Energy Conservation by Sharing the Savings: A Community-Based Approach." Environment and Planning A: Economy and Space 20 (4): 517–34. https://doi.org/10.1068/a200517.

Brown, Marilyn A., and Linda G. Berry. 1995. "Determinants of Program Effectiveness: Results of the National Weatherization Evaluation." Energy 20 (8): 729–43. https://doi.org/10.1016/0360-5442(95)00027-E.

Brown, Marilyn A, and Susan M. Macey. 1985. "Evaluating the Impact of Two Energy Conservation Programmes in a Midwestern City." Applied Geography 5 (1): 39–53. https://doi.org/10.1016/0143-6228(85)90005-0.

Brown, Marilyn A., Dennis L. White, and Steve L. Purucker. 1989. "Energy Savings of Water-Heater Retrofits: Evidence from Hood River." Energy and Buildings 13 (1): 51–61. https://doi.org/10.1016/0378-7788(89)90017-0.

Burgett, Joseph M., and Abdol R. Chini. 2013. "Efficiency Measures for Existing Residential Homes in the Florida Market." International Journal of Construction Education and Research 9, no. 1: 61–80. https://doi.org/10.1080/15578771.2012.677966.

Campbell Systematic Reviews: Policies and Guidelines. Campbell Systematic Reviews 2014: Supplement 1

Clark, Kim E., and David Berry. 1995. "House Characteristics and the Effectiveness of Energy Conservation Measures." Journal of the American Planning Association 61 (3): 386–95. https://doi.org/10.1080/01944369508975649.

Cluett, Rachel, and Jennifer Amann. 2016. "Scaling Up Participation and Savings in Residential Retrofit Programs." American Council for an Energy-Efficient Economy. 39.

Delmas, Magali A., Miriam Fischlein, and Omar I. Asensio. "Information Strategies and Energy Conservation Behavior: A Meta-Analysis of Experimental Studies from 1975 to 2012." *Energy Policy* 61 (October 1, 2013): 729–39. https://doi.org/10.1016/j.enpol.2013.05.109.

Energy Information Administration - EIA. 2019. "EIA Projects Nearly 50% Increase in World Energy Usage by 2050, Led by Growth in Asia - Today in Energy - U.S. Energy Information Administration (EIA)." n.d. Accessed August 13, 2020. https://www.eia.gov/todayinenergy/detail.php?id=41433.

Fels, Margaret F. 1986. "PRISM: An Introduction." Energy and Buildings 9 (1): 5–18. https://doi.org/10.1016/0378-7788(86)90003-4.

Fowlie, Meredith, Michael Greenstone, and Catherine Wolfram. 2018. "Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program *." The Quarterly Journal of Economics 133 (3): 1597–1644. https://doi.org/10.1093/qje/qjy005.

Gillingham, Kenneth, Amelia Keyes, and Karen Palmer. "Advances in Evaluating Energy

Efficiency Policies and Programs.” *Annual Review of Resource Economics* 10, no. 1 (2018): 511–32. https://doi.org/10.1146/annurev-resource-100517-023028.
Gillingham, Kenneth, and Karen Palmer. “Bridging the Energy Efficiency Gap: Policy Insights from Economic Theory and Empirical Evidence.” Review of Environmental Economics and Policy 8, no. 1 (January 1, 2014): 18–38. https://doi.org/10.1093/reep/ret021.
Goldberg, Miriam L. 1986. “A Midwest Low-Income Weatherization Program Seen through PRISM.” Energy and Buildings 9 (1): 37–44. https://doi.org/10.1016/0378-7788(86)90005-8.
Goldman, Charles A., Kathleen M. Greely, and Jeffrey P. Harris. “Retrofit Experience in U.S. Multifamily Buildings: Energy Savings, Costs, and Economics.” Energy 13, no. 11 (November 1, 1988): 797–811. https://doi.org/10.1016/0360-5442(88)90085-0.
Goldman, Charles A., and Ronald L. Ritschard. 1986. “Energy Conservation in Public Housing: A Case Study of the San Francisco Housing Authority.” Energy and Buildings 9 (1–2): 89–98. https://doi.org/10.1016/0378-7788(86)90010-1.
Hartman, Raymond S., and Michael J. Doane.1987. “Taking the Con out of Conservation Program Evaluation.” Resources and Energy 9 (2): 187–207. \ https://doi.org/10.1016/0165-0572(87)90017-X.
Hirst, E., and R. Goeltz. 1985. “Comparison of Actual Energy Savings with Audit Predictions for Homes in the North Central Region of the U.S.A.” Building and Environment 20 (1): 1–6. https://doi.org/10.1016/0360-1323(85)90024-1.
Hirst, Eric. 1985. “Estimating the Long-Term Effects of Utility Energy Conservation Programs: A Pacific Northwest Example.” Technological Forecasting and Social Change 28 (3): 217–29. https://doi.org/10.1016/0040-1625(85)90045-9.
Hirst, Eric. 1987. “Energy and Economic Effects of Utility Financial Incentive Programs: The BPA Residential Weatherization Program.” The Energy Journal 8 (2): 97–110.
Hirst, Eric, Benson Bronfman, Richard Goeltz, John Timble, David Lerman, and Kenneth Keating. 1984. “Evaluation of Utility Residential Energy Conservation Programs: A Pacific Northwest Example.” Energy 9 (3): 193–206. https://doi.org/10.1016/0360-5442(84)90107-5.
Hirst, Eric, and Richard Goeltz. 1985. “Estimating Energy Savings Due to Conservation Programmes: The BPA Residential Weatherization Pilot Programme.” Energy Economics 7 (1): 20–28. https://doi.org/10.1016/0140-9883(85)90035-0.
Hirst, Eric, Richard Goeltz, and David Trumble. 1989. “Effects of the Hood River Conservation Project on Electricity Use.” Energy and Buildings 13 (1): 19–30. https://doi.org/10.1016/0378-7788(89)90014-5.
Hirst, Eric, and David Trumble. 1989. “Effects of the Hood River Conservation Project on Electricity Use and Savings in Single-Family Homes.” Applied Economics 21 (8): 1029. https://doi.org/10.1080/758531240.
Hirst, Eric, Dennis White, and Richard Goeltz. 1985. “Indoor Temperature Changes in Retrofit Homes.” Energy 10 (7): 861–70. https://doi.org/10.1016/0360-5442(85)90119-7.
Hirst, Eric, Dennis White, Richard Goeltz, and Mark Mckinstry. 1985. “Actual Electricity Savings and Audit Predictions for Residential Retrofit in the Pacific Northwest.” Energy and Buildings 8 (2): 83–91. https://doi.org/10.1016/0378-7788(85)90018-0.
IEA. 2019. The Critical Role of Buildings, IEA, Paris https://www.iea.org/reports/the-critical-role-of-buildings
Keating, Kenneth M., and Eric Hirst. 1986. “Advantages and Limits of Longitudinal Evaluation Research in Energy Conservation.” Evaluation and Program Planning 9 (2): 113–20. https://doi.org/10.1016/0149-7189(86)90031-5.
Kerr, Niall, Andy Gouldson, and John Barrett. “The Rationale for Energy Efficiency Policy: Assessing the Recognition of the Multiple Benefits of Energy Efficiency Retrofit Policy.” Energy Policy 106 (July 1, 2017): 212–21. https://doi.org/10.1016/j.enpol.2017.03.053.

Lee, Allen, and Jeffrey Englin. 1989. “Using a Model and Empirical Data to Analyze Manufactured Home Conservation Retrofits.” Energy and Buildings 13 (1): 73–83. https://doi.org/10.1016/0378-7788(89)90019-4.

Liang, Jing, Yueming Qiu, Timothy James, Benjamin L. Ruddell, Michael Dalrymple, Stevan Earl, and Alex Castelazo. 2018. “Do Energy Retrofits Work? Evidence from Commercial and Residential Buildings in Phoenix.” Journal of Environmental Economics and Management 92 (November): 726–43. https://doi.org/10.1016/j.jeem.2017.09.001.

Metcalf, Gilbert E., and Kevin A. Hassett. 1999. “Measuring the Energy Savings from Home Improvement Investments: Evidence from Monthly Billing Data.” The Review of Economics and Statistics 81 (3): 516–28.

Newcomb, Tim M. 1984. “Conservation Program Evaluations: The Control of Self-Selection Bias.” Evaluation Review 8 (3): 425–40. https://doi.org/10.1177/0193841X8400800308.

Rivers, Nicholas, and M. Leslie Shiell. “Free-Riding on Energy Efficiency Subsidies: The Case of Natural Gas Furnaces in Canada.” SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, August 13, 2015. https://doi.org/10.2139/ssrn.2667600.

Rosenstock, S.J. “Issues in Demand-Side Management Programs Operated by Electric Utilities in the United States.” In *IECEC 96. Proceedings of the 31st Intersociety Energy Conversion Engineering Conference*, 3:1598–1606 vol.3, 1996. https://doi.org/10.1109/IECEC.1996.553339.

Scheer, Jim, Matthew Clancy, and Sadhbh Ní Hógáin. 2013. “Quantification of Energy Savings from Ireland’s Home Energy Saving Scheme: An Ex Post Billing Analysis.” Energy Efficiency 6 (1): 35–48. https://doi.org/10.1007/s12053-012-9164-8.

Stern, Paul C., Elliot Aronson, John M. Darley, Daniel H. Hill, Eric Hirst, Willett Kempton, and Thomas J. Wilbanks. 1986. “The Effectiveness of Incentives for Residential Energy Conservation.” Evaluation Review 10 (2): 147–76. https://doi.org/10.1177/0193841X8601000201.

Talwar, Raj, and Eric Hirst. 1981. “Energy Savings from the Minnesota Low-Income Weatherization Programme.” Energy Policy 9 (1): 48–51. https://doi.org/10.1016/0301-4215(81)90206-8.

Tonn, B., E. Rose, and B. Hawkins. 2018. “Evaluation of the U.S. Department of Energy’s Weatherization Assistance Program: Impact Results.” Energy Policy 118 (July): 279–90. https://doi.org/10.1016/j.enpol.2018.03.051.

United Nations Climate Change. 2019. “Climate Ambition Alliance: Nations Renew Their Push to Upscale Action by 2020 and Achieve Net Zero CO2 Emissions by 2050.” Accessed September 8, 2020. https://unfccc.int/news/climate-ambition-alliance-nations-renew-their-push-to-upscale-action-by-2020-and-achieve-net-zero.

United Nations Environment, 2017. Global Status Report. https://www.worldgbc.org/sites/default/files/UNEP%20188_GABC_en%20%28web%29.pdf

Wagner, Barbara Shohl, and Richard C. Diamond. 1987. “The Kansas City Warm Room Project: Economics, Energy Savings and Health and Comfort Impacts.” Energy 12 (6): 447–57. https://doi.org/10.1016/0360-5442(87)90004-1.

Wall, Leonard W., Charles A. Goldman, Arthur H. Rosenfeld, and Gautam S. Dutt. 1983. “Building Energy Use Compilation and Analysis (BECA). Part B: Retrofit of Existing North American Residential Buildings.” Energy and Buildings 5 (3): 151–70. https://doi.org/10.1016/0378-7788(83)90003-8.

Zivin, Joshua Graff, and Kevin Novan. 2016. “Upgrading Efficiency and Behavior: Electricity Savings from Residential Weatherization Programs.” Energy Journal 37 (4): 1–23. https://doi.org/10.5547/01956574.37.4.jziv.

6.0 Appendix

6.1 Sample Selection

Papers that were not accompanied by abstracts were automatically included into the next phase of screening. We decided to include abstracts that met some - but not all - of the inclusion criteria into the next phase of screening to ensure that exclusion decisions were not based on the abstract alone. That said, if the abstract explicitly stated that the study was simulated, located in a place outside of Canada, Europe, or the US, or examined only commercial buildings, it was excluded in this first round of review.

Abstracts that were included into the second round of screening as “maybes” fell into one of six categories: 1) The abstract was unclear about whether or not the model was simulated 2) The abstract discussed energy efficiency policies, “DSM” policies, certificates, and/or subsidies/loans/grants etc. but not necessarily or explicitly home retrofit programs; 3) The abstract was unclear about whether the paper was recommending a methodology for retrofit program assessment, or was actually using the methodology to assess retrofit programs; 4) Limitations discussed in the abstract could affect the causal interpretation; 5) It was unclear from the abstract whether model homes or actual households were used in the study; 6) Outcomes idiosyncratic to the study might have been out of scope of the review (i.e. health outcomes, indoor air pollution outcomes, etc.).

Table A: Keyword Search

Intervention	*Keywords*
Residential retrofit programs	"green home retrofit programs*" OR "residential retrofit*" OR "home retrofit*" OR "home retrofit programs*" OR "weatherization program*" OR "residential retrofit programs*"
Database[1]	*Search results*
SCOPUS (Science Direct/ Elsevier)	485
NBER	5
Web of Science	66
JSTOR	90
SAGE Journals	26
The University of Chicago Press[2]	33
Taylor and Francis	79

TOTAL	784

[1] Databases are listed in the order of which searches were conducted
[2] The University of Chicago Press restricts the keyword search to a limited number of characters, which only permitted the following search term: "residential retrofit programs"

Table B: Inclusion criteria checklist

Author (year)	Text
Title	Text
Name of coder	Text
Study includes applied research (not just theoretical models)	Options: yes/no
Study includes retrofit intervention	Options: yes/no
Study refers to at least one of the outcome variables	Options: yes/no
Study targets private households or individuals in private households	Options: yes/no/discuss
Study was carried out in Europe, Canada, or the US	Options: yes/no
Inclusion decision	Options: include/exclude/
Comments	Comments were used to provide details about papers that the coder was unsure about. Identify why the paper might not be included in the final dataset

6.2 Data Cleaning

In order to calculate mean savings, some data cleaning was required. Some of the studies that evaluated the Bonneville Power Administration (BPA) programs looked at savings one, two, and three years after the initial retrofit was installed; in these cases where the size of the treatment group did not change, savings were averaged over the three years to determine mean savings. Weighted averages were used to determine mean savings in studies where the treatment group changed based on the evaluation period.[9] Three studies about the BPA programs observed savings in two separate treatment groups: households that received an audit and a loan and households that only received an audit (Hirst, 1985; Hirst & Goeltz, 1985; Hirst et al., 1984). This was also observed in Liang et al's (2018) study about the Energize Phoenix Program, which offered three different program structures[10] under its umbrella based on household income, in Brown and White's (1988) study about the Residential Energy Conservation Action Program (RECAP), which stratified savings by program participants located in two different states, and in the Fowlie et al. (2015) and Graff Zivin and Novan (2017) studies, which randomly assigned a behavioural treatment to a subset of their study samples. We considered the savings for each of these groups separately, as per Hoicka et al. (2014), who separated their evaluation of EnerGuide for Houses into four different program iterations based on four changes to the program structure.

6.3 Supplementary Data Tables

Table 3: Household savings by program - weighted average

Name	Savings (%)	Incentive Type	Incentive Amount	Fuel Type	Income
Bonneville Power Administration Pilot	12.7 (n=1409)	Zero Interest Loan	$2200	Electricity	unknown
Bonneville Power Administration Interim	13.2 (n=2207)	Zero Interest Loan	$1330	Electricity	unknown
Energize Phoenix Overall	8 (n=201)	Subsidy	Up to 100% of costs	Electricity	Low-high income
Energize Phoenix Energy Assist 60/40	26 *(n=24)*	Subsidy	60% of costs	Electricity	400% of poverty line
Energize Phoenix Energy Assist 100	0 *(n=14)*	Subsidy	100% of costs	Electricity	200% of poverty line
Energize Phoenix Rebate Match	7 *(n=163)*	Subsidy	Matched utility rebates	Electricity	Higher income
Energy Savings	9.4	Subsidy	$1665 -	Electricity	Low income

[9] For example, if households retrofit in the 1982 iteration of a retrofit program were compared to households retrofit in the 1983 iteration of a retrofit program.

[10] Energize Assist 100 (low income), Energize Assist 60/40 (middle income), and Rebate Match (high income)

Assistance Program	(n=276)		$1735	[11]	
Green Madison & Milwaukee Energy Efficiency	5 (n=79994)	Subsidy and Loan	Subsidy:[12] $1000 - $2500 Loan:[13] $2500 - $25000	Gas and Electricity	Mixed incomes
Home Energy Rebate Offer	1 (n=52)	Cash Rebates	2003 mean: $1300 Up to: $2000	Electricity	Low income
Home Energy Scheme	10.5 (n=210)	Grant	30-35% of costs	Gas	Above low income[14]
Hood River conservation Project	11.5 (n=2428)	Subsidy	100% of costs up to $1.15/ kwh	Electricity	Low income
Low Income Electric Program	20 (n=326)	Subsidy	$1400	Electricity	Low income
Residential Energy Conservation Action Program	5.2 (n=3303)	Subsidy	100% of cost	Electricity	Mixed incomes
SEAI Better Energy Communities	13.8 (n=50)	unknown	unknown	Gas	unknown
San Francisco Housing Authority	8 (n=1822)	Zero Interest Loan	Up to $1000	Gas	Low income
Weatherization Assistance Program	12 (n=39071)	Formula Grant	$650-$6500	Gas and Electric	Low income
Warm Room Retrofit	24.1 (n=5)	Subsidy	100% of cost	Gas	Low income
Total (weighted)	**7.5**				

[11] Zivin and Novan (2016) measured electricity savings, but did not exclude homes in the study based on their primary fuel type
[12] $1000, $1500, and $2000, for a homeowner making investments projected to save 15-24%, 25-34%, or more than 34% of energy
[13] Interest: 4.5% and 5.25% from a local credit union; Amount: $2,500 up to $20,000, up to 100% of installation costs; Terms: 3-10 years
[14] Participants typically need to cover 65–70 % of the cost of measures installed, which potentially crowds-out low income households (Scheer et al., 2013)

	(n= 153,587)	

*Numbers in parentheses indicate the size of the treatment group used to weight savings
**Mean savings were calculated using 29 studies that reported program savings in percent and/or provided a measure of baseline energy consumption from which percent savings could be manually computed

Table 4: Weighted average of savings (%) by population characteristics of homes targeted for retrofits

Savings - Electrically-heated homes	Savings - Fuel-heated homes	Savings - Fuel & electrically-heated homes	Savings - Low-income households	Savings - Middle-high income households	Savings - Undisclosed income
10.1% (n=10,202)	**12.5%** (n=11,771)	**6.8%** (n=131,940)	**12.0%** (n= 43,942)	**5%** (n=2,125)	**12.9%**

Numbers in parentheses denote the number of treated households

Table 5 - Weighted average of savings by study design

Study Design	*N* Measures of Savings	Control Group	Control for Self-Selection	Savings (%)
Simple Difference	8	No	Fixed effects	**12.8** (n= 2,203)
Cross-Section	7	Selected on observables: Eligible non-participants or future program participants	IMR or propensity score matching	**12.2** (n= 3,883)
Difference-in-Difference	10	Selected on observables: Eligible non-participants or future program participants	IMR, propensity score matching, or fixed effects	**11** (n= 14,627)
Randomized Control Trial	3	Randomized treatment and control groups	Controlled for through randomization	**6.8** (n=132,157)

n = number of treated households
Savings are weighted by *n*
4 studies were excluded from these calculations because they did not provide measures of savings in

percent.

Table 6: Measures of cost-effectiveness by program

Program	cost of conserved energy	Payback Period (Years)	Realization Rate (%)
Green Madison and Milwaukee Energy Efficiency		53% of recommended investments do not pay back	58
Bonneville Power Administration Pilot	1981: 2.7 cents/kWh[15] 2020: 7 cents/kWh	10	66
Bonneville Power Administration Interim			1982: 78 1983: 47
Energize Phoenix (overall)	2013: 43.4 cents/kWh[16] 2020: 48 cents/kWh[17]	30.4 - never	28.3
Energy Savings Assistance Program (ESAP)		20	79
Hood River Conservation Project	1984: 0.39 cents/kWh[18] 2020: 0.84 cents/kWh 1986: 8.7 cents/kWh[19] 2020: 19 cents/kWh 1984: 8 cents/kWh[20] 2020: 18 cents/kWh	<1	43
Residential Conservation Service (RCS) Program		<4 to 18	85
San Francisco Housing Authority		3.5	
Weatherization Assistance Program	CO: $6.4/MBtu 1991: 2.18 cents/kWh 2020: 3 cents/kWh	MN, 1977/78: 3.5 CO, Gas: 11 CO, Electric: 3.5 MI, 2011/12: 16	CO: 70 MI, 2011/12: 25

[15] Keating and Hirst, 1986
[16] Liang et al., 2018
[17] Since the retrofits occurred over the period of 2010-2016, we assumed a midpoint of 2013
[18] Brown et al., 1989
[19] Hirst and Trumble, 1989
[20] Hirst et al., 1989

Warm Room Retrofit		2.4-4.6	

Table 7: Net Present Value (NPV)

Program	Net Present Value per household	Net Present Value for the region	Net Present Value for the Program
Bonneville Power Administration Pilot	-$40 to $900[21] $2500 to $2800[22]	$1000 to $2400 per household[23]	-$800 to $2300 per household[24]
Bonneville Power Administration Interim	$1400 to $4600	-$700 to $3200 per household[25]	na
Low Income Electric Program (LIEP)	$485[26]	na	na
San Francisco Housing Authority	$220[27]	$399,000[28]	na
Weatherization Assistance Program	1989: $1660[29]		2008: $1,852 trillion[30] 2010: $5.753 trillion[31]

Table 7: List of Studies

Title	**Author(s)**	**Date**
Measuring the Welfare Effects of Residential Energy Efficiency Programs	Allcott and Greenstone	2017
Actual energy savings after retrofit: Electrically heated homes in the	Hirst	1986

[21] Hirst et al., 1984
[22] Hirst, 1985
[23] Hirst et al., 1984
[24] Hirst et al., 1984
[25] Hirst, 1987
[26] Newcomb, 1984
[27] Goldman and Ritschard, 1986
[28] Goldman and Ritschard, 1986
[29] Brown and Berry, 1995
[30] Tonn et al., 2018
[31] Tonn et al., 2018

Pacific Northwest		
Energy and Economic Effects of Utility Financial Incentive Programs: The BPA Residential Weatherization Program	Hirst	1987
Actual electricity savings and audit predictions for residential retrofit in the pacific northwest	Hirst et al.	1985
Evaluation of utility residential energy conservation programs: A Pacific Northwest example	Hirst et al.	1984
Indoor temperature changes in retrofit homes	Hirst et al.	1985
Estimating energy savings due to conservation programmes: The BPA residential weatherization pilot programme	Hirst and Goeltz	1985
Estimating the long-term effects of utility energy conservation programs: A Pacific Northwest example	Hirst	1985
Energy Conservation for Low-Income Households: The Evaporative Cooler Experience	Ridge	1988
A midwest low-income weatherization program seen through PRISM	Goldberg	1986
Do energy retrofits work? Evidence from commercial and residential buildings in Phoenix	Liang et al.	2018
Upgrading Efficiency and Behavior: Electricity Savings from Residential Weatherization Programs	Graff Zivin and Novan	2016
Evaluating the impact of two energy conservation programmes in a midwestern city	Brown and Macey	1985
Quantification of energy savings from Ireland's Home Energy Saving scheme: an ex post billing analysis	Scheer et al.	2013
Energy savings of water-heater retrofits: Evidence from Hood River	Brown et al.	1989
Effects of the Hood River Conservation Project on electricity use and savings in single-family homes	Hirst and Trumble	1989
Effects of the Hood River Conservation Project on electricity use	Hirst et al.	1989
Using a model and empirical data to analyze manufactured home conservation retrofits	Lee and Englin	1989
Preliminary assessment of the Louisiana Home Energy Rebate Offer program using IPMVP guidelines	Kaiser and Pulsipher	2010
Conservation Program Evaluations: The Control of Self-Selection Bias	Newcomb	1984
Taking the con out of conservation program evaluation	Hartman and Doane	1987
Stimulating Energy Conservation by Sharing the Savings: A Community-Based Approach	Brown and White	1988

Quantitative evaluation of deep retrofitted social housing using metered gas data	Beagon et al.	2018
Advantages and limits of longitudinal evaluation research in energy conservation	Keating and Hirst	1986
Comparison of actual energy savings with audit predictions for homes in the north central region of the U.S.A.	Hirst and Goeltz	1985
Energy conservation in public housing: A case study of the San Francisco housing authority	Goldman and Ritschard	1986
Determinants of program effectiveness: Results of the national weatherization evaluation	Brown and Berry	1995
A utility bill study of mobile home weatherization savings	Burch et al.	1993
Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program	Fowlie et al.	2015
Energy conservation in low-income homes in New York City: the effectiveness of house doctoring	Rodberg	1986
Energy savings from the Minnesota low-income weatherization programme	Talwar and Hirst	1981
Evaluation of the U.S. department of energy's weatherization assistance program: Impact results	Tonn et al.	2018
The Kansas City warm room project: Economics, energy savings and health and comfort impacts	Shohl Wagner and Diamond	1987