N°2025 – 10 • MARCH 2025

WORKING PAPER

Do building energy retrofits deliver savings? A Meta-Analysis

*Marc BAUDRY [1,*2,*3*], Edouard CIVEL[1,4*], Anna CRETI [1,5*], Romain PRESTY [1*,6*]*

Residential energy consumption accounts for approximately 30% of global final energy consumption and 26% of total greenhouse gas emissions. Investing in the energy retrofit of dwellings is widely recognized as a key strategy to reduce both. However, the pace of these investments remains slow, as household decisions are hindered by multiple factors collectively known as the energy efficiency gap. To mitigate this issue, financial support programs for retrofitting have been implemented in developed countries. However, studies assessing their effectiveness reveal significant heterogeneity in outcomes. To contribute to the design of more effective programs, this paper conducts a meta-analysis that clarifies the actual impact of these initiatives, identifies the key determinants of their effectiveness, and corrects potential publication biases in evaluation results. Based on the Levelized Cost of Carbon Abatement derived from the meta-regressions, the main policy implication is that energy retrofitting is cost-effective. However, it should prioritize dwellings using natural gas rather than electricity in Europe, whereas the opposite approach is more suitable for the United States.

JEL CODES: Q48, D61, Q54.

1* Climate Economics Chair – IEF/ILB, France.
2* Economix-CNRS - Paris Nanterre University, France.
3* College of Europe, Bruges, Belgium.
4* Square Research Center – Square Management, France.
5* Paris Dauphine-PSL University, France.
6* IFP School, France.

Acknowledgements : this work has benefited from the financing of the ECOREBA – Building Renovation Economics research initiative of the Climate Economics Research.

Email : marc.baudry@chaireeconomieduclimat.org

KEYWORDS

Energy efficiency | Energy retrofit | Meta-analysis | Meta-regression

Executive summary

Operations of buildings account for 30% of global final energy consumption and 26% of global energy-related Greenhouse Gas (GHG) emissions, 18% if we focus on indirect emissions from the production of electricity and heat used in buildings. Implementing energy efficiency in the building sector helps reduce the energy bill of households on the one hand while addressing climate change on the other hand. However, the weakness of new construction flow compared to the stock explains the relative inertia in energy consumption and GHG emissions. Therefore, **governments in developed countries have designed programs to incentivize investing in energy retrofitting their dwellings by households**. These programs are expected to be win-win to the extent that for one euro invested, at least one euro is saved on the energy bill during the lifetime of the installations.

Given the magnitude of public funds mobilized in energy retrofit programs, empirical works have been conducted to assess their effectiveness. Nevertheless, these works' geographical and temporal scopes are limited due to the cost of implementation and high data requirements, making their comparison and synthesis uneasy. There is, therefore, a **need for a meta-analysis that can summarize the heterogeneous and sometimes contradictory results of existing studies into a synthetic and transferable measure of the effect to be expected from retrofit programs**. This paper is the first to propose such a meta-analysis. The proposed methodology offers several advantages:

1. **Control of heteroscedasticity**: the sampling error varies from one study to another, making it necessary to weight the reported results to reflect their uncertainty.
2. **Control of publication bias**: it is well documented in the literature that studies with high and significant results have a higher chance of being published, inducing a bias in publicly available results.
3. **Control of observed and unobserved heterogeneity**: differences in the characteristics of retrofit programs and methods implemented to assess them have to be neutralized to make reported results comparable.

The meta-regressions implemented on a set of studies reporting up to 171 estimates of the rate of energy savings from retrofit programs conducted all over the world during the four last decades leads to a **significant positive synthetic rate of energy saving of magnitude 10%**, which is nevertheless 2% less than a rough average of results reported in the studies. The meta-regressions results also allow us to compute comparable Levelized Cost of Carbon Abatement (LCCA) from which two crucial policy recommendations can be drawn:

1. It is relevant to prioritize the retrofit of dwellings to abate GHG emissions based on the argument that they are associated with low, or even negative, LCCA.

In Europe, the retrofit of dwellings using natural gas as the main source of energy is more profitable than the retrofit of dwellings using electricity, but it is the opposite in the United States. The sharp difference between Europe and the US stems from the price gap of natural gas for residential use.

1 Introduction

According to the International Energy Agency (IEA), the operations of buildings account for 30% of global final energy consumption and 26% of global energy-related Green House Gas (GHG) emissions, 8% being direct emissions in buildings and 18% indirect emissions from the production of electricity and heat used in buildings[1]. Consequently, implementing energy efficiency in the buildings sector is akin of "killing two birds with one stone". It helps reducing the energy bill of households on the one hand and addressing climate change on the other hand. These two facets of energy efficiency in buildings have long been recognized, as attested for instance by the papers by Rosenfeld et al., 1993 or Levine et al., 1996. However, the weakness of new constructions flow compared to the stock explains the relative inertia in both energy consumption and GHG emissions. In a recent report on energy efficiency (IEA, 2024), including a chapter explicitly dedicated to the buildings sector, the IEA lists a relatively large number of national building energy codes with provisions promoting energy efficiency in new buildings, but notes that significantly fewer include provisions for improving the energy performance of existing buildings. Yet the report points that "improving efficiency in existing buildings is pivotal to accelerate progress, and requires an integrated approach". The question of whether it is relevant to subsidize or assist households in the adoption of retrofit investments, in addition to regulatory measures as those embedded in buildings energy codes, arises.

The argument that a public support to retrofit dwellings is required originates in the gap observed between the expected investment rate in view of the promised profitability and the actual investment rate of retrofit. This energy efficiency gap may be due to households under-valuing the return on investment (Gerarden et al., 2017). As the improvements in buildings' energy efficiency may last for at least 12 years (Jafari and Valentin, 2017), the whole lifetime of these installations may not be properly considered by households. Therefore, governments in developed countries designed multiple policies for households to undertake such investments. These policies are considered win-win to the extent that for one euro invested, at least one euro saved on the energy bill during the lifetime of the installations is expected. The

[1] https://www.iea.org/energy-system/buildings

motivation for this type of public policy has also sometimes been coupled with the promise of positive macroeconomic effects on the economy, particularly in terms of employment (Mikulić et al., 2016). This explains why some programs were launched following major shocks, such as the Weatherization Assistance Program created by the Congress of the United States in 1976 in response to the oil shocks of the 1970s (Tonn et al., 2018) or, in Europe, the requirement for European Union member states to adopt a long-term renovation strategy set out in the Energy Performance of Buildings Directive since 2010, which has been revised in 2018 and recently in 2024, faced with the urgency of acting to limit global warming (Bertoldi et al., 2021) .

The importance of the resources mobilized within the framework of public financial support programs for housing renovation called for the conduct of an evaluation of the effectiveness of these programs. Since the 1980s, a literature specially dedicated to this evaluation has developed. The first evaluation works have immediately highlighted lower results in terms of energy savings actually observed compared to the forecasts made *ex ante* using engineering models. The fine granularity of the data required for the evaluation work, as well as the need to cross-reference measurement data within homes with bill data, local weather data and more general data on the prices of energy supplied to households, have generally limited the geographical scope of each study. Similarly, the cost of monitoring households that have benefited from financial support for renovation has limited the temporal depth of the studies. The literature therefore presents a series of studies with limited geographical and temporal scope. As a consequence, it is difficult to transpose the results to new programs in order to anticipate their effects. Added to this is the evolution and diversity of quantitative methods for evaluating renovation support programs. Hence the need for a meta-analysis which, to our knowledge, has not yet been done.

Meta-analysis refers to the statistical synthesis of results from a series of studies that have been collected systematically, all of these studies aiming to assess the size of a common effect or treatment. As stated in Borenstein et al., 2021:

> If a treatment effect (or effect size) is consistent across the series of studies, these procedures enable us to report that the effect is robust across the kinds of populations sampled, and also to estimate the magnitude of the effect more

> precisely than we could with any of the studies alone. If the treatment effect varies across the series of studies, these procedures enable us to report on the range of effects, and may enable us to identify factors associated with the magnitude of the effect size.

Consistent with this definition, the present paper aims to provide an estimate of the synthetic effect size, in terms of percentage of energy savings, that can be expected from programs supporting the energy retrofit of residential dwellings and to inform on the biases and determinants of energy savings reached. It builds on the now well established methodology of meta-analysis and, more specifically, meta-regressions, the historical development of which is analyzed by Tipton et al., 2018. The idea is to reduce the uncertainty surrounding the impact of such programs, with some studies announcing energy savings of more than 20% (Raynaud et al., 2016, Wagner and Diamond, 1987) while others announce lower results of the order of 1%, potentially not significantly different from 0% (Kaiser and Pulsipher, 2010, Zivin and Novan, 2016). The paper proceeds as follows. A first section presents the meta-data. The choice of the variable of interest, the collection strategy of the studies covered by the meta-analysis and then the general description of these studies are detailed. The second section deals with the publication bias, a bias that appears quite recurrently in meta-analyses and whose correction is often essential to calculate a robust synthetic effect size. A graphical approach completed by a first series of meta-regressions confirms the importance of this bias in our meta-data. A third section follows which aims to control various sources of observed or unobserved heterogeneity between studies. The meta-regressions correcting for these different elements are finally mobilized in order to estimate a robust synthetic effect size around which the uncertainty is greatly reduced with regard to what a rudimentary preliminary examination of the meta-data could suggest. A last section concludes by examining the policy implications of the meta-analysis results.

2 Meta-data

2.1 Collection process

A prerequisite for a meta-analysis and meta-regressions is to identify an outcome of empirical studies that is comparable, computable, reliable and interpretable (Lipsey and Wilson, 2001; Cumpston et al., 2019). It must have the same meaning throughout the different studies and must be convertible in the same unit. This outcome is referred to as the "effect size" estimated by each study. Reliability and comparability requires that a measure of the magnitude of sampling errors is also available, most often in the form of the estimated standard error of the effect size. Looking at the most prominent articles in the field of the economics of energy retrofit programs like the ones by Fowlie et al., 2018 or Webber et al., 2015, it appears that the measure of energy savings post-retrofit, directly or indirectly expressed as a percentage of pre-retrofit energy consumption, is a good candidate. Even if they do not explicitly report such an outcome, many studies display results that enable us to compute this energy savings ratio. For instance, if a study reports the savings in kWh, the total average energy bill in monetary units, and a price per kWh, we are able to compute savings as a percentage of energy consumption. Expressing the outcome as a ratio or percentage eases the comparison by eliminating problems inherent to the use of different units, for instance different currencies if retrofit programs in different countries are studied. Energy savings are a common goal shared by most of energy efficiency retrofit programs due to both the consequences on the energy bill paid by households and the collective consequences in terms of reduction of greenhouse gas emissions or lower investment requirements due to the potential shaving effect on the peak load demand for electricity. As energy savings crucially depend on how much is invested in the retrofit program, an inclusion criteria for studies to be considered in the meta-analysis is that they also display information on the average cost of the program per dwelling and, potentially, information on the decomposition of this cost between different types of actions aimed at improving the energy efficiency of dwellings[2]. In addition to energy

[2]All variables in monetary units have been deflated and converted in 2021 Euros equivalent by using inflation rates and exchange rates data from the OECD. Tests of robustness have been conducting by using the Consumer Price Index instead of the inflation rate and the Purchasing Power Parity instead of the exchange rate.

savings as a percentage of the pre-retrofit energy consumption and to the average cost of the retrofit program per dwelling, we also need to gather information on the retrofit program itself and on the quantitative methods implemented in each study to estimate the effect of the program in terms of energy savings.

With these selection criteria in hand, the next step of a meta analysis is to collect all eligible studies. It can be made by combining two complementary strategies. The first strategy is to build on existing and recent systematic reviews. In our case, we have identified two literature reviews. The first review (Giandomenico et al., 2022) is a census of all studies evaluating energy savings and cost-effectiveness of energy retrofit programs. The second one (Berretta et al., 2021) aims to identify, appraise and synthesize the evidence available on the effectiveness of energy efficiency measure installations, including those bundled with behavioral interventions. The second strategy consists in complementing the set of studies by focusing on papers cited by and citing the most prominent studies identified with the first strategy, namely in our case the two articles by Fowlie et al., 2018 or Webber et al., 2015, including papers published up to October 2023. A check of the relevance of each paper has been made on the basis of its title and its abstract before reading the full paper and searching for the key variables for the meta-regressions. Details on the process and the count of studies collected with these two complementary strategies are given by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) diagram displayed in Figure 1. We ended up with 44 studies, two of which provide two analyses, providing us with 46 different cases. The different studies are listed in Table 1 in chronological order.

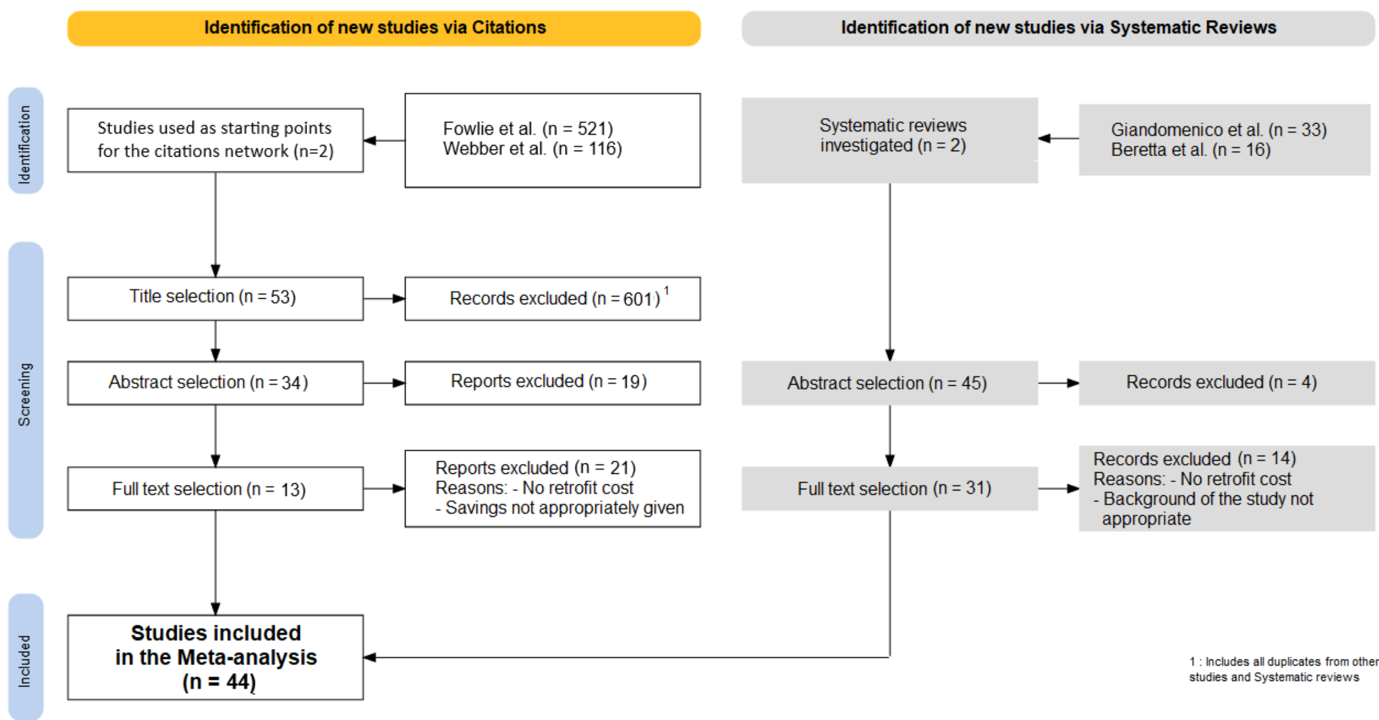
2.2 Overview of studies

The period covered by the studies gathered for our meta-analysis ranges from 1981 to 2022. Nevertheless, Figure 2 highlights that these studies are not evenly distributed over the entire period. They are rather concentrated in two waves. The first wave has spread on the first decade and mainly involves studies at a local scale of the US weatherization assistance program launched in 1976, a detailed presentation of which can be found in Tonn et al.,

Table 1: Studies included in the analysis

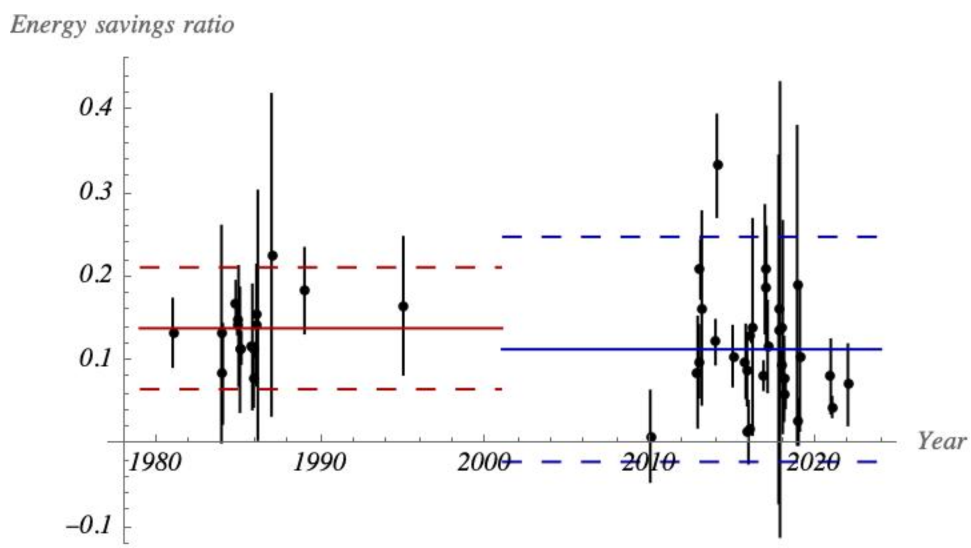
Id	Study	Id	Study
1	Talwar and Hirst, 1981	23	Raynaud, 2014
2	Hirst et al., 1984	24	Webber et al., 2015
3	Newcomb, 1984	25	Zivin and Novan, 2016
4	Hirst et al., 1985a	26	Grimes et al., 2016
5	Hirst et al., 1985b	27	Adan and Fuerst, 2016
6	Hirst, 1985	28	Hamilton et al., 2016
7	Hirst and Goeltz, 1985b	29	James and Ambrose, 2017
8	Hirst and Goeltz, 1985a	30-31	Allcott and Greenstone, 2017
9	Hirst, 1986	32	Coyne et al., 2018
10	Goldberg, 1986	33	Giraudet et al., 2018
11	Keating and Hirst, 1986	34-35	Fowlie et al., 2018
12	Goldman and Ritschard, 1986	36	Liang et al., 2018
13	Rodberg, 1986	37	Beagon et al., 2018
14	Wagner and Diamond, 1987	38	Peñasco and Diaz-Anadon, 2018
15	Hirst and Trumble, 1989	39	Blaise and Glachant, 2019
16	Brown and Berry, 1995	40	Alberini et al., 2019
17	Kaiser and Pulsipher, 2010	41	Filippini and Zhang, 2019
18	Scheer et al., 2013	42	Boampong, 2020
19	Maher, 2013	43	Davis et al., 2020
20	Suter and Shammin, 2013	44	Coyne and Denny, 2021
21	Alberini et al., 2014	45	Peñasco and Anadon, 2021
22	Blasnik et al., 2014	46	Hancevic and Sandoval, 2022

Figure 1: Prisma Diagram



2018. The second wave ranges over the last decade and encompasses geographically more diversified studies, including for instance several European countries like the United Kingdom, Italy, Ireland and France. In Figure 2, the average effect size during the first and second waves is signaled by the solid horizontal line in respectively red and blue, with the associated 95% confidence interval's bounds drawn as dashed lines. Figure 2 exhibits no striking difference in terms of the estimated effect size (points), with an average effect size only slightly lower during the second period. In contrast, the confidence interval sharply widens for the second wave of papers, due to a higher dispersion (vertical segments correspond to 95% confidence intervals individually reported in each study) of estimated effect sizes reported by papers in the second wave compared to papers in the first wave. Moreover, some papers belonging to the second wave do not rule out the hypothesis that there are no significant energy savings, as indicated by the associated confidence interval that includes an energy savings ratio equal to zero. These papers contribute to slightly lowering the average effect size for the second wave and, more importantly, to increase uncertainty surrounding the synthetic effect size as captured by the widened synthetic confidence interval that also includes zero.

Figure 2: Studies by year of publication, with estimated effect size and confidence interval



Examination of papers in the first wave indicates that the main policy stakes they focus on is the reduction of the energy bill paid by households and, more marginally, the shaving of peak load demand when the main source of energy is electricity. In the latter case, the idea is to test whether the weatherization assistance program can avoid having to invest in costly new generation capacities. The focus on the energy bill is consistent with the long lasting impact of the two oil shocks that occurred in 1973 (following the embargo on oil exports to the US and the cuts in oil extraction decided by the Organization of Arab Petroleum Exporting Countries) and 1979 (following a drop in oil production in the wake of the Iranian Revolution). These shocks altered the world oil market and induced a sharp increase in the price of a barrel of oil which has been transmitted to the price of natural gas. A cornerstone of papers in the first wave of studies is the search for an explanation of the gap between energy savings as predicted by engineering models and energy savings actually observed in most *ex post* assessments of the weatherization assistance program conducted for the different US jurisdictions treated in the papers. It led to the emergence of the concept of a rebound effect: improvements in energy efficiency reduce the cost of energy services which can in turn lead to increased energy consumption (Greening et al., 2000). The literature has identified two versions of the rebound effect, the direct rebound effect and the indirect rebound effect. The direct rebound effect can be interpreted as a change in the behavior of households resulting from energy retrofits, meaning that households may adjust their attitudes due to the lower energy use (Belaıd et al., 2018), by increasing their comfort with higher temperatures in the dwelling, by heating rooms that were previously not heated, or by wearing fewer clothes layers on average among other examples. The indirect rebound effect results from households spending money saved thanks to energy efficiency improvements in the purchase of new appliances that use energy. By contrast with the rebound effect which is post-retrofit, a pre-retrofit effect is also mentioned: the pre-bound effect. Indeed, different papers report that old dwellings have an actual energy consumption below their theoretical energy rating (Galvin and Sunikka-Blank, 2016). Overestimating the pre-retrofit consumption therefore leads to overestimating the actual savings post-retrofits.

The renewed interest in the evaluation of energy efficiency programs that gave rise to the second wave of papers seems to stem from the contribution of energy use in buildings to global warming. Indeed, most papers belonging to this second wave point to climate change as an argument to implement such programs. For instance Maher, 2013 starts his introduction by recalling that "buildings account for 42 percent of energy use and 38 percent of CO_2 emissions in the United States". Suter and Shammin, 2013 for their part, start their introduction by stating that "In the face of growing concerns regarding climate change, the extraction and transport of conventional fuels, and energy market volatility, demand-side management through energy efficiency is now at the forefront of energy policy and planning". Similarly, referring to energy efficiency programs based on rebates and tax credits, Alberini et al., 2014 write that "a major goal of these policies is to reduce the emissions of greenhouse gases associated with electricity generation and energy use in the home". The two prominent papers are no exception: Webber et al., 2015 explicitly report results in terms of carbon emissions savings while Fowlie et al., 2018 display estimates for CO_2 abatement costs.

2.3 Main characteristics of studies

Extracting and codifying the information contained in each relevant study on energy savings induced by public policies, more specifically the estimated effect size and its standard deviation, provides with a dataset $\left\{(\hat{\theta}_i, \hat{\sigma}_i, X_i)\right\}_{i=1}^{N}$ where $\hat{\theta}_i$ denotes the estimated effect size in study i, $\hat{\sigma}_i$ stands for the estimated within-study standard deviation of the effect size in study i and X_i is a range of covariates of the effect size. In our specific data analysis, the data set comprises $N = 171$ different estimates of the outcome based on 46 different sub-samples of the whole population of statistical units which are households all over the world, some studies on a same sub-sample of the population providing several estimates from different models. Moreover, we have not been able to retrieve covariates informing about which component of dwellings were subject to retrofit and, more important, about the associated cost per component for each of the 171 different estimates so that we are left with $N = 161$ studies if we are willing to use this kind of information. Thereafter we present results for the three samples referred to as sample (1) with $N = 46$ when only one main estimate per study is used, sample (2) with $N = 171$ when all estimates available in each study are used and sample (3) with $N = 161$ when all estimates in each study are used, conditional on the availability of information on which parts of the dwellings were subject to retrofit and what was the associated cost.

The upper part of Table 2 reports descriptive statistics of the two key elements of the meta analysis, namely the effect size and its standard deviation, which are also illustrated by Figure 2. Overall, the average energy efficiency savings amounts to 12.29% of the initial energy consumption (13.93% for the first wave of studies and 11.41% for the first wave) while the standard deviation equals 3.47% (3.72% and 6.87% for respectively the first wave and the second wave). Hence, energy savings are always significant but the rate of savings is lower and more dispersed across recent studies.

The middle part of Table 2 displays descriptive statitics for covariates used in the meta regressions presented *infra* and related to the retrofit programs assessed by the different studies. These covariates typically inform on the average cost of retrofit in the program, either at the aggregate level (variable "Cost") for samples (1) and (2) or per component of dwellings subject to retrofit for sample (3). In the latter case, the different components considered are walls (variable "Wall") which is ranked fourth in terms of the average cost, attic (variable "Attic") which is ranked second, floor (variable "Floor") which is associated to a low average cost, ducts (variable "Ducts") which is also low in terms of the average cost, heating and cooling system (variable "H/C") which is by far associated to the most expensive average cost, air sealing (variable "Air S.") which is ranked third in terms of the average cost. The last variable "Syst." refers to a retrofit plan to improve the energy efficiency of dwellings viewed as a system, thus selecting one or several components in a comprehensive approach to energy retrofit. It is associated to a low average cost. In addition to these variables expressed in constant 2021 Euros, three other variables characterizing the retrofit programs have been considered. The fist one is a dummy variable indicating whether the retrofit program was more specifically targeting low income households (variable "Low inc."). A little bit less than a third of studies deals with programs that target low income households. The second one is the initial energy consumption (variable "In. Con.") expressed in kWh and the third one is a dummy variable indicating whether natural gas was the main source of energy of dwellings in the study (variable "En. Gas"). depending on the sample considered, more or less half of cases focus on natural gas as the primary source of energy of dwellings.

The lower part of Table 2 relates to the estimation method used in the studies to obtain the effect size. These covariates take the form of dummy variables with value 1 if the method is used in the study and value 0 otherwise. The reference method is a basic Ordinary Least Squares (OLS) regression. It was used by either pioneering works like the one by Hirst et al., 1985a or when it was not possible to adopt another estimation strategy due to data limitations. Among the pioneering works, Hirst et al., 1985a for instance regress actual energy savings observed after a retrofit on characteristics of the dwellings, on the measures undertaken in the context of the weatherization assistance program, on the conservation practices adopted, and on the characteristics of the occupants. This reference method also encompasses early works, the focus of which was to control for weather conditions to make energy consumption before and after the retrofit comparable. In that case, a regression was made of energy consumption on weather conditions before the retrofit and estimation results were then used to compute, for the post retrofit period, what would have been the energy consumption under similar weather conditions without the retrofit (see e.g. Hirst et al., 1985b). The study by Alberini et al., 2014 is illustrative of data constrained regression. Indeed these authors only had access to aggregate energy consumption data before and after the launch of a nation wide tax credit program designed to incentivize Italian homeowners to invest in energy efficiency. Due to the use of structural models (e.g. Allcott and Greenstone, 2017) or the properties of the dependent variable (e.g. Hamilton et al., 2016), some authors rely on alternatives to the OLS estimation method like the maximum likelihood or the simulated moments estimation methods. The corresponding studies are associated to a value 1 for the variable "altOLS" in Table 2. A drawback of basic regression approaches is that households are generally not randomly selected. In a typical retrofit program, a target population is first defined and then members of the targeted population are proposed to enroll in the program if they are eligible. Hence, there is a potential endogeneity of the participation of households or of the amount of financial support provided to retrofit the dwelling. Such an endogeneity occurs if, for instance, those who choose to enroll do it because they anticipate future increased needs of energy due to a change in their own life like the birth of a child or retirement. This risk of endogeneity can be tackled with an instrumental variable (IV) approach as done in early works like the one by Hirst et al., 1984 or lately but with a more elaborated IV strategy by Fowlie et al., 2018. It is referred to as the covariate "IV" in Table 2 and is used in about 10% of studies. Another key issue in the assessment of public programs is to compare the impact on the "treated" group to what happened for a "control" group. Indeed, in the context of retrofit programs for buildings, energy savings for the "treated" group may arise from a change in unobserved variables that actually also had an impact on energy consumption in the "control" group. Matching methods, for instance the Coarsened Exact Matching implemented by Boampong, 2020, help comparing each treated dwelling or household with another one that has not benefited from the program but has similar characteristics. Studies relying on such matching methods are identified by a value 1 of the variable "Matching" in Table 2. Note that, with a matching method, net savings implied by the treatment are not estimated as a coefficient in a regression. An alternative, regression-based, method to compute energy savings net of the variation in the energy bill that also occurred for a control group is the well-known Difference in Differences approach (variable "DiD" in Table 2). About a quarter of the studies rely on a Difference in Differences method. Earlier works like the one by Hirst and Goeltz, 1985b already attempted to compare the change in the energy bill of a "treated" population to that observed for a "control" population without explicitly relying on matching or on the standard DiD regression. The variable "Diff." is then used in Table 2 to refer to this method. Finally, authors have implemented Randomized Controlled Trial (variable "RCT" in Table 2) to avoid self selection bias and endogeneity. In the study by Suter and Shammin, 2013 for instance, students are randomly allocated to different family homes specifically rented to undergraduate students of a same College.

3 Test and correction of publication bias

3.1 General setting

The general specification of a meta analysis may be written as follows

$$\hat{\theta}_i = \theta_i + \beta_0 B_i + \epsilon_i \tag{1}$$

where $\hat{\theta}_i$ stands for the estimated effect size reported in study i and is associated with a sampling error $\epsilon_i \sim N(0, \hat{\sigma}_i^2)$. A peculiarity of meta analysis is the strong suspicion of publication bias which has to be dealt with before to obtain reliable results from the meta regression. One way to correct for the publication bias is to introduce a variable B_i supposed to correctly capture the nature of the bias, as detailed in the next section. A statistically significant coefficient β_0 signals the existence of this type of bias and helps correcting the estimation of the true effect size from this bias. The true effect size θ_i of study i is characterized by

$$\theta_i = \theta + \sum_{j=1}^{J} \beta_j X_{ij} + \mu_i \tag{2}$$

where the X_{ij} $(j \in 1, ..., J)$ are covariates that capture observed heterogeneity between the different studies whereas $\mu_i \sim N(0, \tau^2)$ captures unobserved between-studies heterogeneity and θ is a component of the true effect size which is invariant for all studies. Due to the intrinsic heteroscedasticity induced by sampling errors which are study specific, the linear meta regression defined by (1) and (2) can not be estimated by standard Ordinary Least

Table 2: *Descriptive statitics of studies*

Sample	(1)		(2)		(3)	
Obs	46		171		161	
	Mean	Std	Mean	Std	Mean	Std
$\hat{\theta}_i$ (%)	12.29	6.05	10.60	7.56	10.61	7.51
$\hat{\sigma}_i$ (%)	3.47	2.97	2.68	2.61	2.80	2.65
Cost (constant 2021 €)	3 970	2 777	3 159	2 296	3 199	2 346
Wall (constant 2021 €					398	1281
Attic (constant 2021 €					673	885
Floor (constant 2021 €					66	272
Ducts (constant 2021 €					51	156
H/C (constant 2021 €					1289	1712
Air S. (constant 2021 €					423	812
Syst. (constant 2021 €					66	206
Low inc. (dummy)	0.2826	0.4552	0.3043	0.4615	0.3216	0.4684
In. Con. (kWh)	26 302	16 568	21 309	11 320	21 641	11 984
En. Gas (dummy)	0.4583	0.4261	0.5992	0.4274	0.5809	0.4335
IV (dummy)	0.1304	0.3405	0.0931	0.2914	0.0935	0.2920
Matching (dummy)	0.0434	0.2061	0.0124	0.1111	0.0116	0.1078
DiD (dummy)	0.2173	0.1474	0.2795	0.4501	0.2690	0.4447
altOLS (dummy)	0.0217	0.1474	0.1242	0.3308	0.1169	0.3223
Diff. (dummy)	0.1304	0.3405	0.1863	0.3905	0.1754	0.3814
RCT (dymmy)	0.1739	0.3832	0.1180	0.3236	0.1345	0.3421
Year	2000.79	14.05	2006.23	10.03	2005.85	10.43

Squares (OLS) but either by Weighted Least Squares (WLS) or likelihood based alternative methods as mentioned *infra*.

3.2 Graphical investigation of publication bias

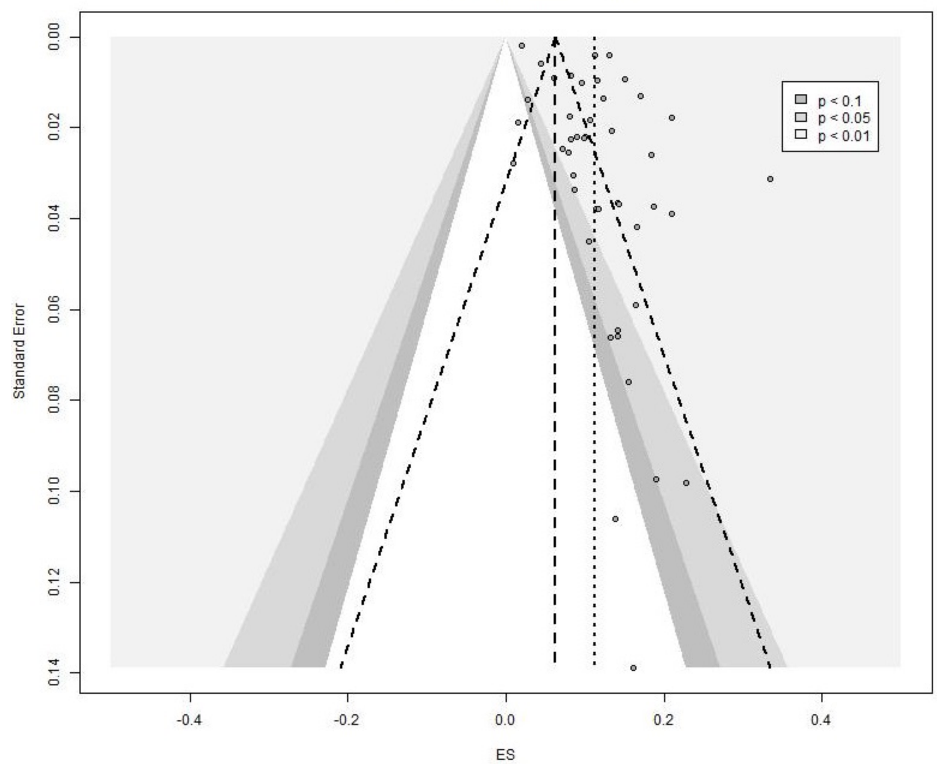
An issue often encountered in meta-analysis is that papers collected for the analysis are most often published papers and that authors may orient results in a specific direction to increase the chance of a paper to be published while editors may prefer to publish papers highlighting significant results. This issue is referred to as the publication bias. Another argument in favor of a risk of publication bias in our specific meta analysis arises from the fact that at least some studies may have been conducted by research centers which have themselves played a role in the inception of public policies to enhance the energy efficiency of buildings. These centers may have been more inclined to highlight evaluations validating the public policies that they have helped to bring about.

It is conventional practice to detect the presence of a publication bias through a funnel plot (T. D. Stanley and Doucouliagos, 2010). Figure 3 displays this funnel plot for our meta-analysis. The areas with different intensity of grey are associated with the usual different p-values for testing whether the result from a given study, as located by its estimated effect size in abscissa and its sampling standard error in ordinates, is significantly different from zero or not[3]. Consistently with Figure 2, Figure 3 highlights that most of studies in our dataset (sample (1)) report a significantly positive effect size, at least when considering 0.1 as the p-value. The thin dashed vertical line positions the average effect size that amounts to 0.123, providing a first but potentially biased estimate of the synthetic effect size and meaning that the energy savings is about 12,3%. In the absence of publication bias, dots should be evenly dispersed on the right and the left of this vertical line, close to it when the standard error is low and more dispersed when the standard error is high. By contrast, a strong asymmetry is observed in Figure 3. Indeed, whereas dots associated with a low standard error are evenly located either on the left side or the right side compared to the average effect size, they all are on the right side when they are associated with a high stan-

[3]This is the "contour enhanced" component of the funnel plot displayed in Figure 3.

dard deviation. This is indicative of a publication bias where only studies obtaining a high estimated effect size are published if the sample size is small and/or the sampling error is high.

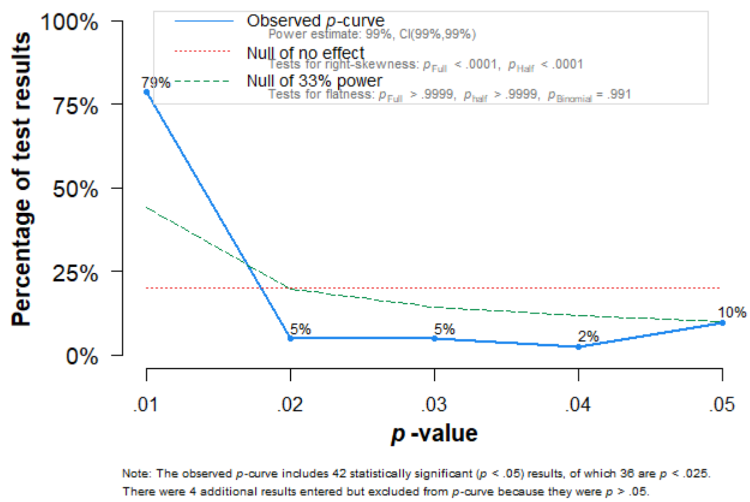
Figure 3: Contour Enhanced Funnel Plot



The asymmetry observed in the funnel plot for high sampling errors suggest to check for the presence of p-hacking. The phenomenon of p-hacking refers to researchers who tweak their analysis until the conventional significance threshold of 0.05 for the p-value of the reported effect size is reached. As suggested by Simonsohn et al., 2014, drawing the p-curve which plots the share of studies obtaining a significant (i.e p-value equal or less than 0.05) estimated effect size as a function of the reported p-value helps detecting p-hacking. Indeed, under the null hypothesis that the true synthetic effect is null the resulting distribution of p-values should be uniform. In case of p-hacking there is an excess of studies with high, though just lower than 0.05, p-values and the distribution is left skewed. Conversely, if the true effect is non-null the distribution should exhibit an excess of low p-values and be right skewed. Figure 4 shows that there is no strong evidence of p-hacking in our set of studies.

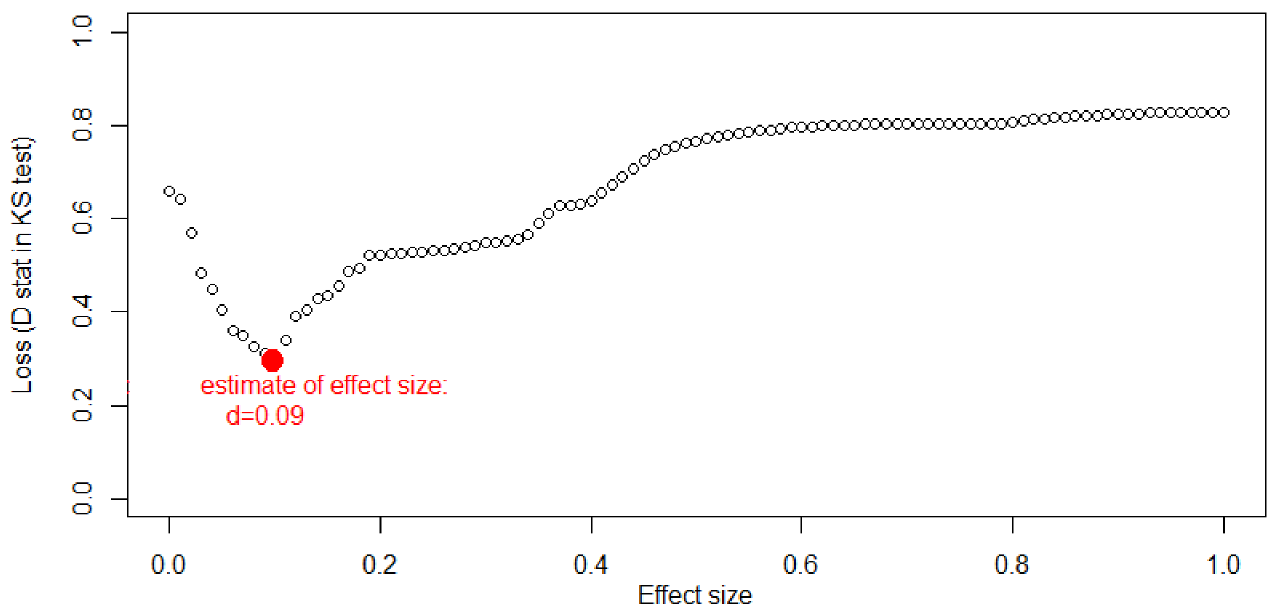
Yet, there is a small excess of studies with a p-value just equal to the 0.05 threshold, which suggests that some studies have been tweaked until the threshold is reached. By contrast, Figure 4 indicates that the null hypothesis of no true effect is very likely rejected due to the sharp right skewness of the distribution of p-values.

Figure 4: P-curve



Simonsohn et al., 2014 proposed to go one step further in the analysis of p-curves and developed the pp-curve which minimum, when a V shaped curve is obtained, provides an estimate of the true synthetic effect size corrected from publication bias. As shown in Figure 5, the resulting synthetic effect size amounts to 9% for our dataset and is 3.3% less than the basic average effect size suggesting that there is an overestimation of energy savings of more or less a third when publication bias is disregarded. Going back to the funnel plot, the thick dashed vertical line in Figure 3 positions the corrected average effect size whereas the cone formed by the two other thick dashed lines indicate the 95% confidence interval for testing whether the different studies significantly depart from the corrected average effect size or not. We find that numerous studies with small sampling error seem to generate significantly higher estimates than the corrected synthetic effect size. However, contrary to the expected shape of a standard funnel plot (T. D. Stanley and Doucouliagos, 2010), the effect sizes reported in studies with low standard errors are abnormally dispersed, which may be indicative of heterogeneity across these studies. Van Aert et al., 2016 found that the PP-curve analysis does not provide a robust estimate of the synthetic effect size when the between study heterogeneity is high, which calls for a specific treatment of heterogeneity as proposed in the next section.

Figure 5: PP-curve



3.3 Regression based tests and correction of publication bias

A commonly used alternative to the pp-curve to correct the synthetic effect size for publication bias, and also a first step towards meta-regressions, is the Precision Effect Test (PET). It relies on the estimation of (1) and (2) under the assumption that there is neither unobserved nor observed between-studies heterogeneity except the one resulting from the publication bias and proxied by the standard error (T. D. Stanley, 2008, T. D. Stanley and Doucouliagos, 2014). Said another way, PET posits that $B_i = \sigma_i$ in (1) and thus consists in estimating the relation

$$\theta_i = \theta + \beta_0 \sigma_i + \epsilon_i \tag{3}$$

In (3), heteroscedasticity stems from $\epsilon_i \sim N(0, \hat{\sigma}_i^2)$. It can be tackled either by dividing both sides of (3) by σ_i or equivalently by using a Weighted Least Square (WLS) estimation method with weights $1/\sigma_i$. Egger et al., 1997 proposed the conventional t-test of $\beta_0 = 0$ as a test of publication bias, but highlighted that it has low power. The test is actually also a Funnel Asymmetry Test (FAT) to the extent that studies with a larger standard deviation report a larger value of the estimated effect size. Conversely, the t-test for the intercept θ is a powerful test for the existence of a genuine synthetic effect size beyond publication selection. Estimation results reported in Table 3 advocate in favor of a positive genuine synthetic effect size in spite of a highly significant publication bias whatever the sample considered. According to the PET method, energy saving would nevertheless amount to 4.46% with sample (1) or 2.28% and 2.27% with samples (2) and (3), which is three times to six times lower than the rough estimate based on the average of reported effect sizes and also less than the corrected synthetic effect size computed with the PP-value approach. Table 3 complements the picture by providing an alternative to the basic average of reported effect sizes as a benchmark. This alternative is the fixed effect model corresponding to the estimation of (1) and (2) under the assumption that there is no source of heterogeneity and no publication bias. It amounts to computing the weighted average of reported effect sizes, with the inverse of the sampling variances of studies as weights. Interestingly, the estimated synthetic effect size with this fix effect model is close to the one obtained with the PP-curve approach. Thereafter, the fix effect model is referred to as Model A and the more elaborated model that accounts for a publication bias as described in (3) is referred to as Model B.

PET is attractive and extensively used due to its simplicity, but the method admits some drawbacks. In the presence of publication bias, only significant effect sizes are reported, hence the sampling errors are drawn from a truncated distribution. Rosenberger and T. Stanley, 2009 explain that the reported effect size and its standard deviation then do not have a linear relationship but a non linear relationship that stems from the truncation. T. D. Stanley and Doucouliagos, 2014 simulations show that this non linearity is satisfactorily treated by setting $B_i = \sigma_i^2$ instead of $B_i = \sigma_i$ in (1). This results in the Precision Effect Estimate with Standard Error (PEESE) associated to the following regression:

$$\theta_i = \theta + \beta_0 \sigma_i^2 + \epsilon_i \tag{4}$$

Heteroscedasticity in (4) is treated as in (3). As for the PET method, testing for the significance of the slope and the intercept respectively reveals the existence of a publication bias and of a genuine non zero synthetic effect size. Table 3 shows that PEESE results are consistent with that of PET. The synthetic effect size corrected for publication bias with PEESE indicates that energy saving amounts to 6.05% with sample (1) and 4.25% and 4.26% with samples (2) and (3), which is higher than with PET but is still half the basic average of reported effect sizes and less than the corrected synthetic effect size computed with the PP-value approach or the estimated synthetic effect size with the fix effect model. Nevertheless, in the absence of control of between studies heterogeneity other than the one captured by the standard deviation or its squared value, PET and PEESE models may be misspecified as discussed in the next section.

Table 3: Fixed effect model without and with control of publication bias

Sample	(1)	(2)	(3)
Obs	46	171	161
	Model A (*Fixed effect*)		
θ	0.095179***	0.071857***	0.071613***
	Model B (*PET* version)		
θ	0.044639***	0.028039***	0.027307***
β_0	3.277422***	4.126395***	4.357351***
Adj R^2	0.1995	0.19	0.1968
	Model B (*PEESE* version)		
θ	0.060572***	0.042748***	0.042679***
β_0	30.069906	42.591316***	45.373273***
Adj R^2	0.03598	0.04252	0.04254
Note:			*p<0.1; **p<0.05; ***p<0.01

4 Meta-regressions

4.1 Observed between-studies heterogeneity

The funnel plot displayed in Figure 3 exhibits an abnormal dispersion of effect sizes reported for studies with low standard deviation. Indeed, when their standard deviation is low, studies should report effect sizes close to the common synthetic effect size θ according to (3) and (4), and thus also close to each others. By contrast, a high dispersion indicates that these studies do not share a common true effect size due to some sources of heterogeneity. Part of the heterogeneity is actually observed and can be directly controlled for, the remaining heterogeneity having to be dealt with by introducing a random effect.

Meta-regressions allow to control for observed between-studies heterogeneity. Indeed heterogeneity arises from the fact that the sample of individuals are not the same or the implementation of the different studies has not been strictly identical. Typically, the total budget devoted to the retrofit of dwellings of programs evaluated by the different studies are not the same, as indicated by the relatively high standard deviation of the "Cost" covariate in Table 2, leading to a different amount of investment per dwelling and different degrees of retrofit. To control for this major source of heterogeneity in a meta regression it was key collecting only studies which explicitly report the amount of investment per unit of dwelling. The period covered by the studies included in our meta-analysis also span several decades and the technical solutions involved in retrofit programs may have evolved and improved in terms of cost-efficiency, resulting in higher energy savings. As all studies do not display information on all potential sources of heterogeneity, there is a trade-off between, on the one hand, collecting studies with information available on many potential sources at the cost of ending up with a small dataset of studies and, on the other hand, disregarding some potential sources to prioritize the dataset size when information is not provided often enough. The trade off led to the set of covariates reported in Table 2. Moreover, given the significant publication bias outlined in the previous section, we kept on correcting it in the following PET-like meta-regression that accounts for observed heterogeneity through the vector X_i of values for study i of the different covariates

$$\theta_i = \theta + \beta_0 \sigma_i + \sum_{j=1}^{J} \beta_j X_{ij} + \epsilon_i \tag{5}$$

In parallel, we also considered a PEESE-like meta-regression to address the truncation arising from the selection of only studies with a significant effect size. The associated estimated equation is

$$\theta_i = \theta + \beta_0 \sigma_i^2 + \sum_{j=1}^{J} \beta_j X_{ij} + \epsilon_i \tag{6}$$

Being subject to the same heteroscedasticity than (3) and (4), models in (5) and (6) have also to be estimated by WLS. With the aim to identify the contribution of covariates to observed hetereogeneity, we proceed in two steps.

In a first step we only consider covariates related to the context of the different studies, also named substantive variables. The corresponding estimation results are reported in Table 4 and are referred to as results for Model C, . As expected, the cost of the retrofit has a highly significant positive impact on the energy savings. The reported coefficients for the "Cost" variable for samples (1) and (2) and its decompositions in different elements for sample (3) are quasi elasticities. Hence, they directly inform on the percentage of energy savings per constant 2021 Euro invested in retrofit, all other things being equal. For instance, every thousand constant 2021 Euros invested in retrofit is estimated to have resulted in 1.29% of energy savings according to estimation results of the PET version of the model for sample (1). The decomposition of the total cost in its different components as done in sample (3) indicates that retrofit acting on walls, attic, floor and heating/cooling system has a significant positive impact on energy savings whereas no significant impact is obtained for action on ducts, air sealing and on the whole system. These results are robust with respect to the sample of studies used and whether the PET or PEESE version of the model is considered. Similarly, a robust result is obtained as regards the higher energy savings resulting from the retrofit of dwellings the main source of energy of which is natural gas. By contrast, energy savings are not significantly sensitive to the initial consumption in regressions on sample (3) whereas they are when using samples (1) and (2). The focus made by a retrofit program on low income households does not have a clear cut impact on energy savings, regressions on samples (2) and (3) suggesting a significant positive impact that does not appear when using sample (1).

Table 4: Meta regression (Model C)

	PET			PEESE		
Sample	(1)	(2)	(3)	(1)	(2)	(3)
Obs	46	171	161	46	171	161
θ	-0.053083***	-0.028533***	-0.002468	-0.050779***	-0.028292***	-0.000229
β_0	1.063492	1.596146***	0.732057	9.310312	17,537984	9.312729
Cost	1.29E-05***	9.29E-06***		1.25E-05***	9.48E-06***	
Wall			3.88E-05***			3.90E-05***
Attic			4.83E-06**			4.47E-06**
Floor			1.01E-04***			1.01E-04***
Ducts			7.91E-06			1.87E-05
H/C			1.37E-05***			1.39E-05***
Air S.			1.25E-05			1.36E-05
Syst.			3.68E-05			4.64E-05
Low inc.	0,003256	0.065319***	0.088330***	0.009813	0.068604***	0.090435***
In. Con.	3.15E-06***	9.42E-07**	-5.49E-07	3.41E-06***	1.27E-06***	-5,05E-07
En. Gas	0.026249*	0.047259***	0.038820***	0.026247*	0.048790***	0.039057***
Adj R^2	0.6981	0.5329	0.6207	0.6858	0.5195	0.6194
Note:	*p<0.1; **p<0.05; ***p<0.01					

In a second step we add covariates related to the estimation method used in the different studies. It yields estimation results for Model D reported in Table 5. This second step follows on from a focus made in the literature on meta-regressions on the risk of confounding between methodological factors and other substantively interesting moderators describing the context of the studies. Durlak and Lipsey, 1991 and later Lipsey, 2003 provide examples of how methodological differences in studies can lead to misinterpretation of the impact of substantive variables and argue in favor of the inclusion of methodological variables in addition to substantive variables in meta-regressions to correctly tackle the risk of confounding. Such a confounding problem would arise if results reported for substantive variables in Table 5 were drastically different from those reported in Table 4. It is actually not the case. The main results stressed for Model C remain valid for Model D, except for the magnitude of the "Cost" covariate if sample (2) is used and the significance and/or magnitude of coefficients associated to the different cost components when the regression is conducted on sample (3). The other results that depart in Model D from Model C are limited to substantive variables that already had no clear cut impact in Model C (e.g. variables "Low inc." and "In. Con."). Finally, consistent with the absence of a strong confounding problem, only a few of the methodological variables have a significant impact in Model D and, when they do have one, the significance of the impact varies depending on the sample under consideration.

4.2 Unobserved between-studies heterogeneity

In addition to observed heterogeneity, the different studies collected for a meta-analysis also entail unobserved sources of heterogeneity, for instance differences in the details of implementation of each study. In that case, it is generally assumed that the "true" effect sizes of the different studies are drawn from a same normal distribution. Combined with the existence of observed heterogeneity and with the PET/PEESE correction for publication bias, this random effect yields the general model setting described by (1) and (2) or, equivalently, by a more condensed form that expresses the observed effect size of a study as the sum of four terms:

$$\hat{\theta}_i = (\theta + \sum_{j=1}^{J} \beta_j X_{ij}) + \beta_0 B_i + \mu_i + \epsilon_i \tag{7}$$

The first term in brackets yields the expected effect size for study i, conditional on the substantive and methodological variables characterizing that study. The second term captures the publication bias with $B_i = \sigma_i$ in the PET version of the meta-regression and $B_i = \sigma_i^2$ in the PEESE version. The last term captures the sampling error of study i. Finally $\mu_i \sim N(0, \tau^2)$ captures the unobserved between-studies heterogeneity. The total variance of $\hat{\theta}_i$ is now $\hat{\sigma_i}^2 + \tau^2$. WLS can no longer be used in that kind of meta-regression because τ is unknown and has to be estimated jointly with the other parameters. A Restricted Maximum Likelihood

Table 5: Meta regression (Model D)

	PET			*PEESE*		
Sample	(1)	(2)	(3)	(1)	(2)	(3)
Obs	46	171	161	46	171	161
θ	2.549232*	1.754380	1.083312	2.758120*	2.353606	1.303242
β_0	1.432695**	1.921848***	0.799912	13.226047	18.923832*	10.051915
Cost	1.18E-05***	7.57E-06***		1.23E-05***	7.30E-06***	
Wall			4.86E-05***			4.90E-05***
Attic			2.88E-06			2.48E-06
Floor			5.76E-05**			5.61E-05**
Ducts			2.81E-05			3.90E-05
H/C			1.35E-05***			1.37E-05***
Air S.			2.09E-05**			2.18E-05**
Syst.			3.41E-06			9.11E-06
Low inc.	0.037477*	0.015306	0.044349*	0.046357*	0.024114	0.047421**
In. Con.	2.92E-07	4.06E-07	-7.99E-07	4.91E-07	7.85E-07	-7.21E-07
En. Gas	0.043004**	0.049561***	0.038868**	0.047680*	0.052398***	0.039528***
IV	-0.053042**	-0.007894	-0.031305	-0.058617**	-0.012119	-0.031814
Matching	0.040596	0.053256	0.070539	0.051206	0.065097	0.077020
DiD	-0.032128*	0.003519	0.007953	-0.039696**	-0.005976	0.006453
GLM	0.021019	-0.008429	-0.012735	0.008338	-0.021803	-0.015811
Diff.	0.007692	0.072411***	0.060046**	-0.002772	0.054751**	0.056312**
RCT	-0.023010	0.004641	0.014953	-0.026898	0.008405	0.016286
Year	-0.001269*	-0.000884	-0.00053	-0.001369*	-0.001176	-0.000647
Adj R^2	0.8383	0.5808	0.6613	0.8212	0.5657	0.6602

Note: *p<0.1; **p<0.05; ***p<0.01

(RML) method can then be used. This estimation method fixes the problem of high-dimensional data caused by a small number of observations and a high number of variables. It provides unbiased estimates of the magnitude of unobserved heterogeneity (Langan et al., 2019). Estimation results for this so-called mixed effects model, thereafter referred to as Model E, are reported in Table 6.

Estimation results seem to be slightly sensitive to the introduction of the random effect capturing unobserved heterogeneity. The magnitude and the ranking of the quasi elasticities associated to expenditures in the different retrofit components when estimating the model on sample (3) are affected, as well as the magnitude of the total cost when using sample (2). The only cost component with a non significant impact on energy savings is that associated to ducts retrofit. It is nonetheless consistent with the non significance obtained in Models C and D. A noticeable difference compared to Models C and D is also that a retrofit treating the dwelling as a whole system (substantive variable "Syst.") is now ranked first in terms of energy savings per invested Euro. Methodological variables do not significantly affect estimated energy savings, except for the "Diff." method which was already found to be the most impactfull method in Model D. But, before all, the systematically highly significant estimated value of parameter τ strongly supports the rationale for estimating the mixed effects model. It reveals that there is indeed unobserved heterogeneity at play in our meta-analysis and that accounting for it is key to provide reliable results as regards the synthetic effect size. It is also noteworthy that coefficient β_0 measuring publication bias is systematically significant and positive whatever the sample used and the choice between the PET and PEESE control method, whereas it was not the case in Models C and D, more specifically in the PEESE version of the meta-regression. This result suggests that the publication bias is probably not correctly controlled for if unobserved heterogeneity is disregarded.

4.3 Estimates and distribution of the synthetic effect size

In view of the previous results and in order to produce a synthetic effect size that best summarizes the results of the different studies, this section proposes to reason on a study that is itself synthetic. Indeed, it appears necessary to neutralize the strong heterogeneity

Table 6: Meta regression (Model E)

	PET			PEESE		
Sample	(1)	(2)	(3)	(1)	(2)	(3)
Obs	46	171	161	46	171	161
θ	0.116293	1.276990	−0.807678	0.252075	−1.229777	−0.475342
β_0	0.984533**	1.196596***	1.086513***	8.606994*	11.523741***	10.198118***
Cost	1.33E-05***	1.25E-05***		1.39E-05***	1.30E-05***	
Wall			1.52E-05***			1.67E-05***
Attic			1.44E-05***			1.26E-05***
Floor			4.11E-05**			3.97E-05**
Ducts			3.16E-05			4.04E-05
H/C			1.20E-05***			1.29E-05***
Air S.			1.28E-05**			1.40E-05**
Syst.			5.74E-05**			6.19E-05**
Low inc.	0.028030	0.012928	0.018528	0.032638*	0.018675	0.024127**
In. Con.	1.24E-06	1.51E-06***	1.57E-06***	1.41E-06*	1.72E-06***	1.71E-06
En. Gas	0.027745	0.047003***	0.051828***	0.027067	0.044494***	0.050084***
IV	−0.044485*	−0.013957	−0.018979	−0.047188*	−0.020142	−0.022375
Matching	0.040120	0.054356*	0.058734*	0.043303	0.056954*	0.063336**
DiD	0.000516	0.010322	0.007897	0.000629	0.006282	0.006171
GLM	0.019137	−0.018892	−0.020864	0.007205	−0.026793*	−0.027074**
Diff.	0.01097	0.063498***	0.050923***	0.002413	0.054503***	0.044712***
RCT	−0.032307	−0.012672	−0.011595	−0.032380	−0.009414	−0.004640
Year	−6.28E-05	0.000616	0.000380	−0.000125	0.000599	0.000220
τ	0.037906***	0.039073***	0.036529***	0.038350***	0.039774***	0.036974***
Adj R^2	0.3145	0.6456	0.6814	0.2983	0.6328	0.6736

Note: *p<0.1; **p<0.05; ***p<0.01

observed between studies. To do this, this section considers a fictitious study having the average characteristics of the studies used for each sample. By substituting the parameters estimated using meta-regression (7) for model E and the values of the variables at the average point, a synthetic effect size associated with the study that is itself synthetic can be calculated.

The evaluation at the mean point neutralizes the impact of the heterogeneity observed between studies. However, the resulting synthetic effect size is itself only an estimate, the precision of which has to be assessed. The uncertainty around this estimate more specifically follows on from that of the estimated parameters whose synthetic effect size is a linear combination. It is evaluated by using Monte Carlo simulations of the vector of coefficients involved in (7) drawn from a multinormal distribution whose expectation vector is the vector of estimated parameters and the variance-covariance matrix is that of the estimated coefficients resulting from the meta-regression. 10 000 random draws have been carried out, from which a 95% confidence interval has then been computed for the synthetic effect size at the mean point of the observations. The effect of unobserved heterogeneity can be isolated by comparing, on the one hand, the synthetic effect size obtained using 10 000 random draws of the random term μ_i knowing the estimated value of its standard deviation τ (Model E bis in Table 7, Table 8 and Table 9), and on the other hand, the synthetic effect size obtained by assuming that the random term μ_i is equal to zero (Model E in Table 7, Table 8 and Table 9).

It may be informative to isolate the impact of publication bias as well. In the basic PET and PEESE models as presented in (3) and (4), it is generally considered that the "true" synthetic effect size is given by the intercept, whereas the publication bias is evaluated by the additional term $\beta_0\sigma_i$ or $\beta_0\sigma_i^2$. We follow this logic by comparing the estimates of the mean effect size and its confidence interval obtained in two ways. The first way is to assume that σ_i (PET version) or σ_i^2 (PEESE version) takes the mean value over the sample, like the other variables. It corresponds to the estimates and confidence intervals in the columns headed "No" in Table 7, Table 8 and Table 9 for the adjustment of the publication bias. The second method is to assume that the sampling error measure of the synthetic study is zero. It corresponds to the estimates and confidence intervals in the columns headed "Yes" in Table 7, Table 8 and Table 9 for the adjustment of the publication bias. It is important to emphasize that the "No" and "Yes" columns only differ by the value given to the sampling error for the synthetic study, the estimated coefficients being those obtained by controlling for the possible publication bias in the meta-regressions.

Table 7: Synthetic effect size with sample (1)

Method		PET		PEESE	
Adjustment for the publication bias		No	Yes	No	Yes
Model A	$\bar{\theta}$	0.095		0.095	
	CI	[0.078 , 0.112]		[0.078 , 0.112]	
Model B	$\bar{\theta}$	0.158	0.044	0.096	0.060
	CI	[0.092 , 0.224]	[0.027 , 0.061]	[0.050 , 0.143]	[0.044 , 0.075]
Model C	$\bar{\theta}$	0.130	0.094	0.114	0.103
	CI	[0.095 , 0.166]	[0.074 , 0.114]	[0.089 , 0.139]	[0.088 , 0.118]
Model D	$\bar{\theta}$	0.129	0.079	0.111	0.095
	CI	[0.103 , 0.155]	[0.059 , 0.100]	[0.092 , 0.130]	[0.080 , 0.110]
Model E	$\bar{\theta}$	0.123	0.089	0.115	0.105
	CI	[0.107 , 0.139]	[0.064 , 0.114]	[0.101 , 0.129]	[0.088 , 0.121]
Model E bis	$\bar{\theta}$	0.123	0.088	0.115	0.105
	CI	[0.046 , 0.199]	[0.009 , 0.168]	[0.039 , 0.191]	[0.029 , 0.182]

The process described above to compute a synthetic effect size and to assess its sensitivity to the publication bias and to unobserved heterogeneity is replicated for the three samples of cases. Table 7, Table 8 and Table 9 present the estimated synthetic effect size and confidence interval for respectively sample (1) sample (2) and sample (3). A noticeable feature is that the confidence interval never includes zero. Hence, all our estimates of the synthetic effect size attest the existence of positive and significant energy savings resulting from retrofit programs. The confidence interval also never includes the basic average effect size reported in Table 2 (which is 0.1229 for sample (1), 0.106 for sample (2) and 0.1061 for sample (3)) when there is adjustment for the publication bias, except for Model E bis which accounts for unobserved heterogeneity. When outside the confidence interval, the basic average effect size is systematically higher than the upper bound of the confidence interval. In contrast, in the absence of adjustment of the synthetic effect size for the publication bias, the basic average effect size is most often included in the confidence interval of the synthetic effect size, except for Model A whatever the sample, and some other Models depending on the sample considered. The basic average of reported effect sizes of the studies thus encompasses a publication bias which makes it significantly overstate the estimated "true" synthetic effect size. It is only when accounting for the high unobserved heterogeneity between studies (Model E bis) that the publication bias between the basic average effect size and the estimated "true" synthetic effect is no longer significant.

Figure 6 to Figure 9 complement the analysis of the synthetic effect size by comparing the box and whisker plot of its distribution generated with Monte Carlo simulations of the different models and that generated with similar simulations for a basic model based on the average value and variance of observed effect sizes for sample (1). The box and whisker plots highlight that the range of values of the estimated "true" synthetic effect size is narrower than the range of values of the "basic" approach to the effect size. Meta-regressions thus appear to generate a more precise summary of energy savings reported in the surveyed studies, whatever the model and the sample used, even when adding uncertainty resulting from unobserved between studies heterogeneity (Model E bis). The comparison between Figure 6 and Figure 7 confirms the impact of the publication bias for the PET version of meta-regressions: adjusting for the publication bias results in a lower and narrower dispersion of simulated values of the synthetic effect size, at least for Models B C and D. Similar results are obtained for the PEESE version of the meta-regressions on the basis of the comparison between Figure 8 and Figure 9.

Table 8: Synthetic effect size with sample (2)

Method		PET		PEESE	
Adjustment for the publication bias		No	Yes	No	Yes
Model A	$\bar{\theta}$	0.071		0.071	
	CI	[0.062 , 0.081]		[0.062 , 0.081]	
Model B	$\bar{\theta}$	0.144	0.028	0.076	0.042
	CI	[0.107 , 0.181]	[0.019 , 0.036]	[0.052 , 0.100]	[0.035 , 0.049]
Model C	$\bar{\theta}$	0.114	0.070	0.093	0.080
	CI	[0.089 , 0.139]	[0.058 , 0.082]	[0.077 , 0.110]	[0.071 , 0.088]
Model D	$\bar{\theta}$	0.114	0.060	0.093	0.078
	CI	[0.090 , 0.138]	[0.041 , 0.079]	[0.076 , 0.109]	[0.064 , 0.091]
Model E	$\bar{\theta}$	0.108	0.081	0.098	0.089
	CI	[0.100 , 0.116]	[0.068 , 0.093]	[0.091 , 0.105]	[0.080 , 0.097]
Model E bis	$\bar{\theta}$	0.109	0.080	0.098	0.088
	CI	[0.032 , 0.185]	[0.002 , 0.158]	[0.020 , 0.177]	[0.010 , 0.166]

Table 9: *Synthetic effect size with sample (3)*

Method		PET		PEESE	
Adjustment for the publication bias		No	Yes	No	Yes
Model A	$\bar{\theta}$	0.071		0.071	
	CI	[0.061 , 0.081]		[0.061 , 0.081]	
Model B	$\bar{\theta}$	0.144	0.027	0.075	0.042
	CI	[0.107 , 0.181]	[0.018 , 0.035]	[0.051 , 0.098]	[0.035 , 0.050]
Model C	$\bar{\theta}$	0.107	0.087	0.099	0.093
	CI	[0.083 , 0.130]	[0.070 , 0.104]	[0.085 , 0.114]	[0.081 , 0.104]
Model D	$\bar{\theta}$	0.106	0.084	0.099	0.092
	CI	[0.083 , 0.129]	[0.064 , 0.105]	[0.084 , 0.113]	[0.079 , 0.104]
Model E	$\bar{\theta}$	0.106	0.077	0.099	0.091
	CI	[0.099 , 0.114]	[0.065 , 0.089]	[0.092 , 0.106]	[0.083 , 0.099]
Model E bis	$\bar{\theta}$	0.106	0.077	0.099	0.091
	CI	[0.034 , 0.179]	[0.004 , 0.150]	[0.026 , 0.173]	[0.018 , 0.164]

Figure 6: Box and Whisker Plot of the distribution of the synthetic effect size: PET model without adjustment for the publication bias

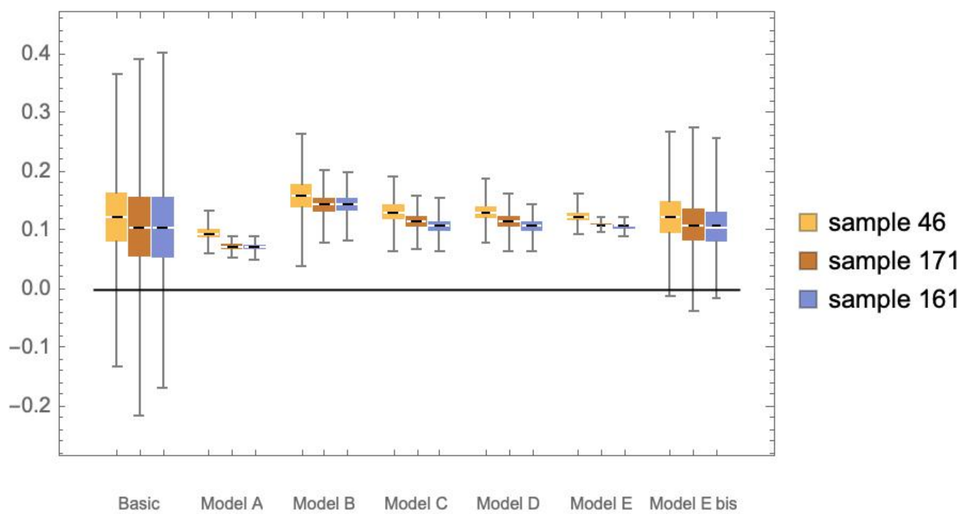


Figure 7: Box and Whisker Plot of the distribution of the synthetic effect size: PET model with adjustment for the publication bias

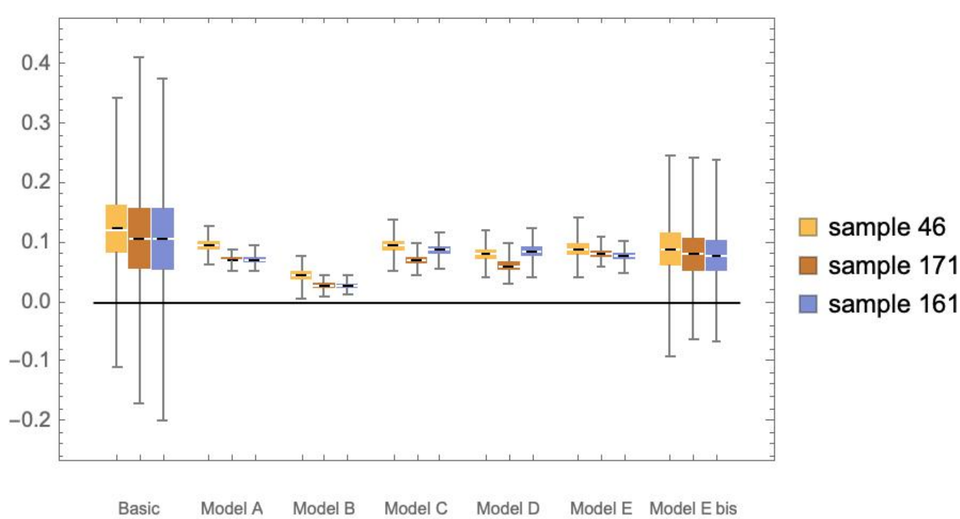


Figure 8: Box and Whisker Plot of the distribution of the synthetic effect size: PEESE model without adjustment for the publication bias

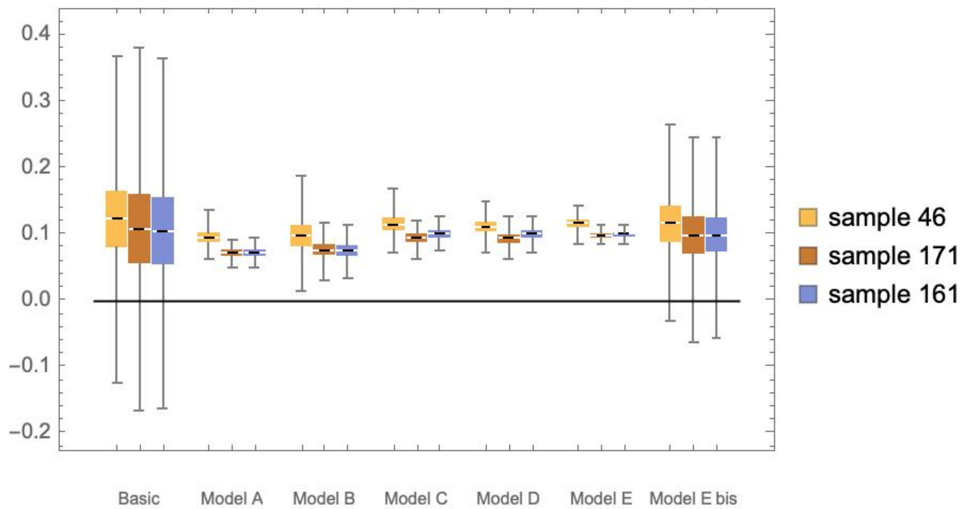
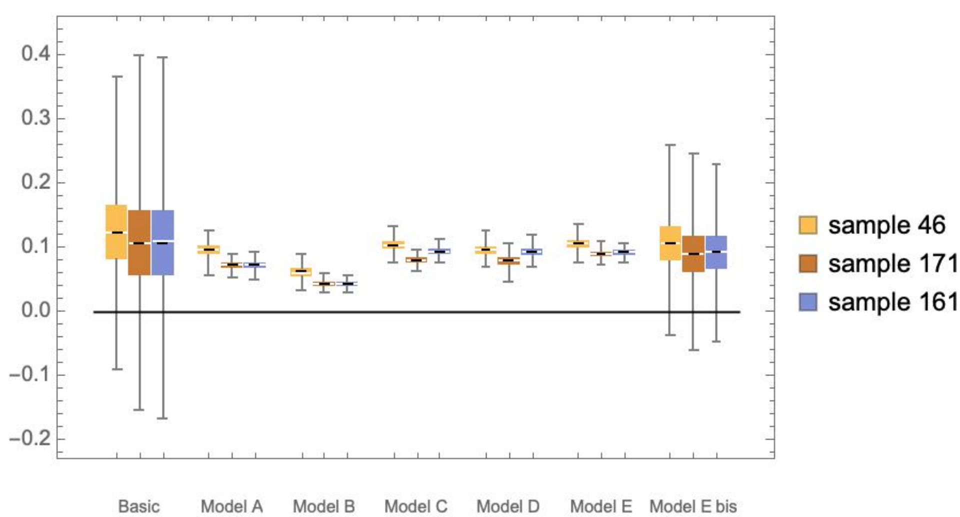


Figure 9: Box and Whisker Plot of the distribution of the synthetic effect size: PEESE model with adjustment for the publication bias

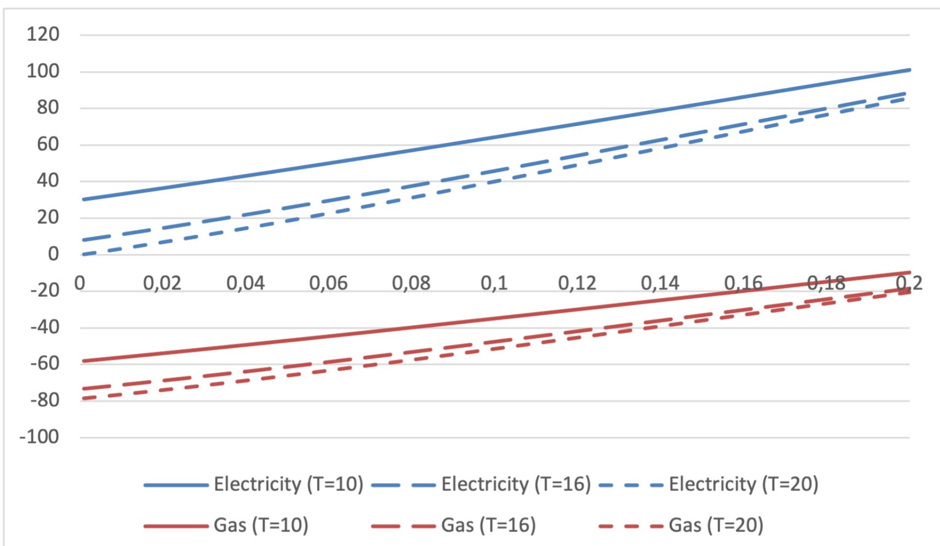


5 Conclusion and policy implications

The meta-analysis carried out in this paper confirms the importance of correcting for publication bias on the one hand, and of taking into account the heterogeneity of studies, including and especially unobserved heterogeneity, on the other hand. The correction of the publication bias leads to reducing the energy savings rate by at most 2% compared to a basic average of the rates reported in the studies. In return, and under the joint effect of taking into account the observed or unobserved heterogeneity and a correct treatment of heteroscedasticity, a substantial reduction in the uncertainty around the measurement of the synthetic energy savings rate is obtained. In view of the initial uncertainty surrounding this measurement, the gain is appreciable and offers greater visibility as to the effects to be expected from public policies promoting the retrofitting of household housing. Based on the mean point of the observations for the largest sample, knowing that the observed and estimated values differ little when considering a sample enlarged to different versions of the results per study, and using the PEESE version of the most complete meta-regression model with correction of the publication bias, the estimated energy saving rate is 6.31% and amounts to an annual saving of 1 345 kWh if an energy other than natural gas is used by the dwelling. The estimated energy saving rate increases to 10.76% if natural gas is the main energy source, i.e. an annual saving of 2 293 kWh. All of these figures result from an average retrofit cost of 3 159 Euros or, at the average exchange rate for the year 2023, 3 490 US dollars.

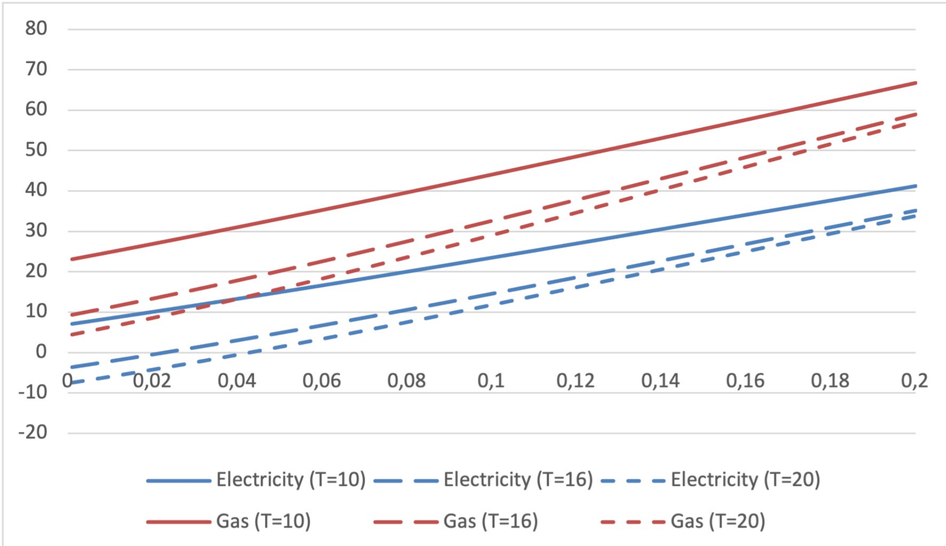
A first lesson in terms of policy recommendation, consistent with the motivations of the retrofit programs evaluated during the first decade of the period covered by our meta-analysis, concerns the profitability of the investments thus made. To do this, even with a back of the envelope calculation, it is necessary to know the price of energy in order to calculate the bill reduction. As this calculation depends on the energy source, the geographical area, and the year considered, we consider four cases by crossing the case of gas *versus* electricity and the case of the European Union *versus* the United States for the year 2023. Let first consider the case of electricity as the primary source of energy for dwellings. According to Eurostat, the average price within the European Union was 0.1125€/kWh in 2023. It results in yearly energy savings of 151.36€. According to the US Energy Information Agency, the average price of electricity for residential end use was 0.1648$/kWh and results in yearly energy savings of 221.73$. When natural gas is considered as the primary source of energy, its average price for residential consumers amounts to 0.2937€/kWh within the European Union according to Eurostat and to a much lower value of 0.0509$/MWh in the US according to the US Energy Information Agency. As a consequence, we compute that yearly energy savings are 673.64€ in the European Union and 119.03$ in the US. Retrofit of dwellings with gas as the primary source of energy is thus much more attractive in the European Union compared to the US, due to gap in the price of gas for residential use between the two sides of the Atlantic. Note also that retrofit of dwellings using gas is more profitable than retrofit using electricity in Europe and conversely in the US.

Figure 10: Sensitivity of the Levelized Cost of Carbon Abatament (€/t) in Europe with respect to the discount rate (abscissa) and the project lifespan (T, in years)



A second lesson in terms of policy recommendations can be drawn, inspired by the focus of the second wave of studies mostly concentrated on the last decade of the period covered by our meta-analysis: is retrofit of dwellings a low cost strategy to reduce Green House Gas emissions? To address this question, we need to complement our back of the envelope calculation with data on electricity carbon intensity and carbon intensity of natural gas burned as a fuel. Electricity carbon intensity in 2023 is assumed to be 0.00021 tonnes of equivalent CO_2 per kWh in Europe according to Eurostat and 0.00394 tonnes of equivalent

Figure 11: Sensitivity of the Levelized Cost of Carbon Abatement (€/t) in the United States with respect to the discount rate (abscissa) and the project lifespan (T, in years)



CO_2 per kWh in the United States according to the US Energy Information Administration, the difference stemming from the electricity mix. We assess the profitability of retrofit with the Levelized Cost of Carbon Abatement (LCCA) proposed by Baker and Khatami, 2019. The LCCA is the threshold value of the price of a tonne of equivalent CO_2 that makes a low carbon project just profitable according to the Net Present Value criteria. As a result, the LCCA can be computed as the ratio between the retrofit cost net of the sum of discounted yearly energy savings in the numerator and the discounted sum of avoided tonnes of equivalent CO_2 in the denominator. Where there is pricing of GHG emissions, either in the form of a carbon tax or in the form of a price on an Emission Trading System, a LCCA lower than the level of pricing implies that the low carbon project is profitable and conversely. Prior computing the LCCA, we have to set the lifespan of the project and the discount rate. In order to make our calculation comparable to that of Fowlie et al., 2018, we consider a lifespan of 16 years and a 3% interest rate. It follows on that the LCCA of the retrofit of a dwelling using electricity as the main source of energy amounts to 18.03€/t in Europe whereas it is 1.19€/t in the US. If we compare it to the price of a tonne of carbon on the EU-ETS that was higher than or close to 80€/t all along year 2023, we conclude that retrofit programs are clearly profitable in Europe and even more in the US. When it comes to natural gas as the main source of energy for dwellings, the carbon intensity of gas burned as a fuel has the same value of 0.00018088675 tonnes of equivalent CO_2 per kWh in Europe and the US. It follows on that the LCCA is -66.38€/t in Europe and 15.52€/t in the US. The negative LCCA in Europe is due to energy savings offsetting the gross cost of the retrofit, hence the negative numerator. It is consistent with the Enkvist et al., 2007 abatement cost of emissions in buildings but sharply contrasts with the high positive net cost obtained by Fowlie et al., 2018. Actually the rate of energy savings obtained by the latter is similar to that estimated in our meta-analysis, but Fowlie et al., 2018 stress the particularly, not to say abnormal, high average cost of retrofit in their study which lead to a positive and significantly higher LCCA for the US. The sharp difference between the highly negative LCCA in Europe and the positive LCCA in the US stems from the price gap of natural gas for residential use in the two geographical zones. A sensitivity analysis of the results with respect to the discount factor and the lifespan of the project reveals that our main results as regards the LCCA are robust, as illustrated by Figure 10 and Figure 11. We conclude from these calculations that prioritizing the retrofit of dwellings to abate GHG emissions based on the argument that there are associated to low, or even negative, net abatement cost is relevant. Nevertheless, the magnitude of the profitability as measured by the LCCA of such low carbon investments crucially depends on what is the main source of energy used in these dwellings and what is the price of this source of energy in the geographical zone under consideration.

References

Adan H. and Fuerst F. (2016). "Do energy efficiency measures really reduce household energy consumption? A difference-in-difference analysis". *Energy Efficiency* 9.5, pp. 1207–1219.

Alberini A., Bigano A., and Boeri M. (2014). "Looking for free riding: energy efficiency incentives and Italian homeowners". *Energy Efficiency* 7.4, pp. 571–590.

Alberini A., Khymych O., and Scasny M. (2019). "The Elusive Effects of Residential Energy Efficiency Improvements: Evidence from Ukraine".

Allcott H. and Greenstone M. (2017). *Measuring the welfare effects of residential energy efficiency programs.* Tech. rep. National Bureau of Economic Research.

Baker E. D. and Khatami S. N. (2019). "The levelized cost of carbon: a practical, if imperfect, method to compare CO2 abatement projects". *Climate Policy* 19.9, pp. 1132–1143.

Beagon P., Boland F., and O'Donnell J. (2018). "Quantitative evaluation of deep retrofitted social housing using metered gas data". *Energy and Buildings* 170, pp. 242–256.

Belaıd F., Bakaloglou S., and Roubaud D. (2018). "Direct rebound effect of residential gas demand: Empirical evidence from France". *Energy policy* 115, pp. 23–31.

Berretta M., Furgeson J., Zamawe C., Hamilton I., Wu Y., Ferraro P. J., Haddaway N., and Eyers J. (2021). "PROTOCOL: Residential energy efficiency interventions: An effectiveness systematic review". *Campbell Systematic Reviews* 17.4, e1205.

Bertoldi P., Economidou M., Palermo V., Boza-Kiss B., and Todeschi V. (2021). "How to finance energy renovation of residential buildings: Review of current and emerging financing instruments in the EU". *Wiley Interdisciplinary Reviews: Energy and Environment* 10.1, e384.

Blaise G. and Glachant M. (2019). "Quel est l'impact des travaux de rénovation énergétique des logements sur la consommation d'énergie". *La revue de l'énergie* 646, pp. 46–60.

Blasnik M., Dalhoff G., Pigg S., Mendyk A., Carroll D., and Ucar F. (2014). "National Weatherization Assistance Program Impact Evaluation: Energy Impacts for Single Family Homes". *ORNL/TM-2015/13, Oak Ridge National Laboratory, Oak Ridge, Tennessee, September*.

Boampong R. (2020). "Evaluating the energy-saving effects of a utility demand-side management program: A difference-in-difference coarsened exact matching approach". *The Energy Journal* 41.4, pp. 185–208.

Borenstein M., Hedges L. V., Higgins J. P., and Rothstein H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Brown M. A. and Berry L. G. (1995). "Determinants of program effectiveness: Results of the national weatherization evaluation". *Energy* 20.8, pp. 729–743.

Coyne B. and Denny E. (2021). "Retrofit effectiveness: Evidence from a nationwide residential energy efficiency programme". *Energy Policy* 159, p. 112576.

Coyne B., Lyons S., and McCoy D. (2018). "The effects of home energy efficiency upgrades on social housing tenants: evidence from Ireland". *Energy Efficiency* 11.8, pp. 2077–2100.

Cumpston M., Li T., Page M. J., Chandler J., Welch V. A., Higgins J. P., and Thomas J. (2019). "Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions". *Cochrane Database Syst Rev* 10.10.1002, p. 14651858.

Davis L. W., Martinez S., and Taboada B. (2020). "How effective is energy-efficient housing? Evidence from a field trial in Mexico". *Journal of Development Economics* 143, p. 102390.

Durlak J. A. and Lipsey M. W. (1991). "A practitioner's guide to meta-analysis". *American Journal of community psychology* 19, pp. 291–332.

Egger M., Smith G. D., Schneider M., and Minder C. (1997). "Bias in meta-analysis detected by a simple, graphical test". *Bmj* 315.7109, pp. 629–634.

Enkvist P., Nauclér T., and Rosander J. (2007). "A cost curve for greenhouse gas reduction". *McKinsey Quarterly* 1, pp. 35–45.

Filippini M. and Zhang L. (2019). "Impacts of heat metering and efficiency retrofit policy on residential energy consumption in China". *Environmental Economics and Policy Studies* 21.2, pp. 203–216.

Fowlie M., Greenstone M., and Wolfram C. (2018). "Do energy efficiency investments deliver? Evidence from the weatherization assistance program". *The Quarterly Journal of Economics* 133.3, pp. 1597–1644.

Galvin R. and Sunikka-Blank M. (2016). "Quantification of (p)rebound effects in retrofit policies–Why does it matter?" *Energy* 95, pp. 415–424.

Gerarden T. D., Newell R. G., and Stavins R. N. (2017). "Assessing the energy-efficiency gap". *Journal of Economic Literature* 55.4, pp. 1486–1525.

Giandomenico L., Papineau M., and Rivers N. (2022). "A systematic review of energy efficiency home retrofit evaluation studies". *Annual Review of Resource Economics* 14.1, pp. 689–708.

Giraudet L.-G., Houde S., and Maher J. (2018). "Moral hazard and the energy efficiency gap: Theory and evidence". *Journal of the Association of Environmental and Resource Economists* 5.4, pp. 755–790.

Goldberg M. L. (1986). "A Midwest low-income weatherization program seen through PRISM". *Energy and buildings* 9.1-2, pp. 37–44.

Goldman C. A. and Ritschard R. L. (1986). "Energy conservation in public housing: A case study of the San Francisco housing authority". *Energy and buildings* 9.1-2, pp. 89–98.

Greening L. A., Greene D. L., and Difiglio C. (2000). "Energy efficiency and consumption—the rebound effect—a survey". *Energy policy* 28.6-7, pp. 389–401.

Grimes A., Preval N., Young C., Arnold R., Denne T., Howden-Chapman P., and Telfar-Barnard L. (2016). "Does retrofitted insulation reduce household energy use? Theory and practice". *The Energy Journal* 37.4.

Hamilton I. G., Summerfield A. J., Shipworth D., Steadman J. P., Oreszczyn T., and Lowe R. J. (2016). "Energy efficiency uptake and energy savings in English houses: A cohort study". *Energy and Buildings* 118, pp. 259–276.

Hancevic P. I. and Sandoval H. H. (2022). "Low-income energy efficiency programs and energy consumption". *Journal of Environmental Economics and Management* 113, p. 102656.

Hirst E. and Goeltz R. (1985a). "Comparison of actual energy savings with audit predictions for homes in the north central region of the USA". *Building and Environment* 20.1, pp. 1–6.

Hirst E. (1985). "Estimating the long-term effects of utility energy conservation programs: A Pacific Northwest example". *Technological Forecasting and Social Change* 28.3, pp. 217–229.

– (1986). "Actual energy savings after retrofit: Electrically heated homes in the Pacific Northwest". *Energy* 11.3, pp. 299–308.

Hirst E., Bronfman B., Goeltz R., Timble J., Lerman D., and Keating K. (1984). "Evaluation of utility residential energy conservation programs: a Pacific Northwest example". *Energy* 9.3, pp. 193–206.

Hirst E. and Goeltz R. (1985b). "Estimating energy savings due to conservation programmes: The BPA residential weatherization pilot programme". *Energy Economics* 7.1, pp. 20–28.

Hirst E. and Trumble D. (1989). “Effects of the Hood River Conservation Project on electricity use and savings in single-family homes”. *Applied Economics* 21.8, pp. 1029–1042.

Hirst E., White D., and Goeltz R. (1985a). “Indoor temperature changes in retrofit homes”. *Energy* 10.7, pp. 861–870.

Hirst E., White D., Goeltz R., and Mckinstry M. (1985b). “Actual electricity savings and audit predictions for residential retrofit in the Pacific northwest”. *Energy and buildings* 8.2, pp. 83–91.

IEA (2024). *Energy Efficiency 2024 (https://www.iea.org/reports/energy-efficiency-2024)*. International Energy Agency.

Jafari A. and Valentin V. (2017). “An optimization framework for building energy retrofits decision-making”. *Building and environment* 115, pp. 118–129.

James M. and Ambrose M. (2017). “Retrofit or behaviour change? Which has the greater impact on energy consumption in low income households?” *Procedia Engineering* 180, pp. 1558–1567.

Kaiser M. J. and Pulsipher A. G. (2010). “Preliminary assessment of the Louisiana Home Energy Rebate Offer program using IPMVP guidelines”. *Applied Energy* 87.2, pp. 691–702.

Keating K. M. and Hirst E. (1986). “Advantages and limits of longitudinal evaluation research in energy conservation”. *Evaluation and Program Planning* 9.2, pp. 113–120.

Langan D., Higgins J. P., Jackson D., Bowden J., Veroniki A. A., Kontopantelis E., Viechtbauer W., and Simmonds M. (2019). “A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses”. *Research synthesis methods* 10.1, pp. 83–98.

Levine M. D., Price L., and Martin N. (1996). “Mitigation options for carbon dioxide emissions from buildings: A global analysis”. *Energy Policy* 24.10-11, pp. 937–949.

Liang J., Qiu Y., James T., Ruddell B. L., Dalrymple M., Earl S., and Castelazo A. (2018). “Do energy retrofits work? Evidence from commercial and residential buildings in Phoenix”. *Journal of Environmental Economics and Management* 92, pp. 726–743.

Lipsey M. W. (2003). “Those confounded moderators in meta-analysis: Good, bad, and ugly”. *The Annals of the American Academy of Political and Social Science* 587.1, pp. 69–81.

Lipsey M. W. and Wilson D. B. (2001). *Practical meta-analysis*. Vol. 49. Sage Publications Thousand Oaks, CA.

Maher J. (2013). “Evaluating the Cost-Effectiveness of Rebate Programs for Residential Energy-Efficiency Retrofits”.

Mikulić D., Bakarić I. R., and Slijepčević S. (2016). “The economic impact of energy saving retrofits of residential and public buildings in Croatia”. *Energy Policy* 96, pp. 630–644.

Newcomb T. M. (1984). “Conservation Program Evaluations: The Control of Self-Selection Bias”. *Evaluation Review* 8.3, pp. 425–440.

Peñasco C. and Anadon L. D. (2021). “Assessing the effectiveness of energy efficiency measures in the residential sector through dynamic treatment effects: Evidence for the United Kingdom”. *Energy, COVID, and Climate Change, 1st IAEE Online Conference, June 7-9, 2021*. International Association for Energy Economics.

Peñasco C. and Diaz-Anadon L. (2018). “Energy efficiency measures in the residential sector: Implementing mechanisms for successful low-carbon transitions in households”.

Raynaud M. (2014). “Evaluation ex-post de l’efficacité de solutions de rénovation énergétique en résidentiel”. PhD thesis. Paris, ENMP.

Raynaud M., Osso D., Bourges B., Duplessis B., and Adnot J. (2016). “Evidence of an indirect rebound effect with reversible heat pumps: having air conditioning but not using it?” *Energy efficiency* 9, pp. 847–860.

Rodberg L. S. (1986). “Energy conservation in low-income homes in New York City: the effectiveness of house doctoring”. *Energy and buildings* 9.1-2, pp. 55–64.

Rosenberger R. S. and Stanley T. (2009). “Publication selection of recreation demand price elasticities: a meta-analysis”. *Work. Pap., Oregon State Univ./Hendrix Coll.*

Rosenfeld A., Atkinson C., Koomey J., Meier A., Mowris R. J., and Price L. (1993). “Conserved energy supply curves for US buildings”. *Contemporary Economic Policy* 11.1, pp. 45–68.

Scheer J., Clancy M., and Hógáin S. N. (2013). “Quantification of energy savings from Ireland’s Home Energy Saving scheme: an ex post billing analysis”. *Energy Efficiency* 6.1, pp. 35–48.

Simonsohn U., Nelson L. D., and Simmons J. P. (2014). “P-curve: a key to the file-drawer.” *Journal of experimental psychology: General* 143.2, p. 534.

Stanley T. D. (2008). “Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection”. *Oxford Bulletin of Economics and statistics* 70.1, pp. 103–127.

Stanley T. D. and Doucouliagos H. (2010). “Picture this: a simple graph that reveals much ado about research”. *Journal of Economic Surveys* 24.1, pp. 170–191.

– (2014). “Meta-regression approximations to reduce publication selection bias”. *Research Synthesis Methods* 5.1, pp. 60–78.

Suter J. F. and Shammin M. R. (2013). “Returns to residential energy efficiency and conservation measures: A field experiment”. *Energy Policy* 59, pp. 551–561.

Talwar R. and Hirst E. (1981). “Energy savings from the Minnesota low-income weatherization programme”. *Energy Policy* 9.1, pp. 48–51.

Tipton E., Pustejovsky J., and Ahmadi H. (2018). “A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018.” *Research Synthesis Method* 10.2, pp. 161–179.

Tonn B., Rose E., and Hawkins B. (2018). “Evaluation of the US department of energy’s weatherization assistance program: Impact results”. *Energy Policy* 118, pp. 279–290.

Van Aert R. C., Wicherts J. M., and Assen M. A. van (2016). “Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve”. *Perspectives on Psychological Science* 11.5, pp. 713–729.

Wagner B. S. and Diamond R. C. (1987). “The Kansas City warm room project: Economics, energy savings and health and comfort impacts”. *Energy* 12.6, pp. 447–457.

Webber P., Gouldson A., and Kerr N. (2015). “The impacts of household retrofit and domestic energy efficiency schemes: A large scale, ex post evaluation”. *Energy Policy* 84, pp. 35–43.

Zivin J. G. and Novan K. (2016). “Upgrading efficiency and behavior: electricity savings from residential weatherization programs”. *The Energy Journal* 37.4.

WORKING PAPER

PREVIOUS ISSUES

Working Paper Publication Directors :

Marc Baudry, Philippe Delacote, Olivier Massol