Do S&L programmes deserve a 'robust evaluation' label?

John Fawcett, Databuild Research & Solutions (UK)

ABSTRACT

Standards & Labelling programmes are believed to represent a cost effective way to overcome market failures in the sale of energy efficient appliances and equipment. The focus in relation to S&L programmes is often upon MVE, establishing compliance with a standard or label. Less clear is the difference the existence of the standard or label makes to the energy using product market, and so net energy savings.

Post-implementation studies in this regard are not common and understandably tend to include caveats around the accuracy of the 'business as usual' scenario and the extent of influence of wider policy and landscape changes. This can result in arbitrary impact allocation to S&L programmes simply because their implementation coincides with noticeable market or product changes. Yet there remain key questions around:

- Market direction – it is often not clear how far the standard or label has led businesses to do things they would not otherwise have done, as oppose to businesses simply achieving recognition for something they had already been intending to do / were doing.
- Customer response - studies have queried the extent of evidence bases on customer recognition, comprehension and preferences for energy labels. This would therefore affect label or standard influence on sales, business perceptions of customer requirements and so influence on market transformation.

This paper - built upon literature reviews and interviews with key stakeholders and evaluators - aims to understand:

- How the programmes are perceived by managers - to propel market transformation or simply as a reward for those businesses looking to innovate and improve their products? – and what effect this has upon the way they are evaluated.
- How far robust attribution methodologies (e.g. consideration of counterfactual scenarios) are deployed in programme impact assessments or evaluations.
- Therefore how far we can be confident in the claimed impacts and benefits S&L programmes deliver.
- What additional evaluation would be useful, in what way, and what the perceived challenges are to doing this.

The findings generated will enable an assessment of what steps could enhance the evaluation of these programmes and to what extent the challenges to accurately understanding their benefits can be overcome.

Standards and Labelling programmes

Although often packaged in a single term, and with the same overarching purpose of market transformation, there are three distinct types of S&L programme, with quite differing specific objectives:

Type of programme	Description	Intention
Minimum Energy Performance Standards (MEPS)	An energy efficiency performance standard for a particular product, intended to provide a minimum baseline that all products in that group must meet in order to be allowed to be sold to consumers.	'Push' effect; seeks to force manufacturers / retailers to improve the efficiency of their products and effectively cuts the lowest performing out of the market.
Compulsory labelling 	Labelling - that all products within a certain group need to carry - which illustrates the level of energy performance (generally exemplified by star rating labels – *see left*). This enables consumers to make informed choices.	'Push' and 'pull' effects; participation is compulsory but can also be an opportunity for the best performing (though value is reliant upon consumer awareness of – and importance attached to – the labels).
Endorsement labelling 	Labels that can be applied to the best performing products within a particular group to indicate this to consumers.	'Pull' effect; no compulsion but provides an incentive for industry to develop products (though again value is dependent upon consumer engagement).

In many countries, the three types of programme are deployed concurrently as part of a suite of measures designed to influence market transformation and deliver energy / carbon savings.

S&L programme evaluation

At a basic level, assessment of the effectiveness of S&L programmes requires evaluators to establish what a product market looks like now (in terms of sales and market share of the compliant / most efficient products) compared to pre-intervention.

MEPS evaluation, or at least attribution of market effects, is *in principle* clear-cut. MEPS effectively deem certain products to be non-compliant and they are removed from the market, therefore any post-implementation effects and consumer purchasing shifts can be easily correlated with the timing of the policy coming into effect (and / or manufacturers and retailers adapting in advance). The least efficient products disappear and – assuming similar levels of sales - more efficient products replace them.

There are differences between evaluations for compulsory and endorsement labelling (not least that the former tends to be much more resource intensive). However, overall both types of labelling evaluations, again *in principle*, observe effects since introduction upon the sales and market share of the most efficient / endorsed products.

It is relatively straightforward to observe the efficiency of a product group at point A and the efficiency at Point B and see that there has been substantial change. The challenges

arise in being able to attribute the observed changes in consumption to programs implemented and determine what how much of this would have happened anyway.

S&L evaluation challenges

Timing of evaluation

Impact assessments conducted prior to the introduction of MEPS and labelling programmes are essential as an evidence base / business case needs to be built to demonstrate the need for – and potential value of – the programme.

Post-implementation evaluations and impact assessments of S&L policies do take place, as well as very empirical approaches e.g. annual in-store shelf-stock surveys to estimate changes in apparent compliance with standards and labelling.

However, in many cases, evaluation prior to implementation has been the *only* study conducted regarding the label. By definition, such studies cannot make use of any data outside of pilots / lab testing, which limits their reliability.

Considering the counterfactual

"Baseline development is often highly contentious and, at best, a good guess of what might have been." Vine et al (2000)

Regardless of whether evaluation is conducted before or after the introduction of the S&L programme, to gain an accurate sense of the effect of the S&L programme, evaluators need to build an accurate baseline for comparison. This generally takes one of two forms:

1. Comparison of a treatment group to a control group, which is similar in all respects except the variable / condition being assessed. In the context of S&L programmes, this would mean comparing to a territory very similar to that introducing the S&L programme. Ehler and Talerico (1999) highlighted that selection of an appropriate comparison territory would need to consider a range of factors, including economic performance, climate, population, educational levels, workforce, existing policy and programme framework, product sales, household income and poverty, energy consumption per capita, and homeownership rates. On this basis, finding a suitable control can be very challenging and it is likely there is no very suitable match. In addition, collection of similar market data in the treatment region *and* the control/comparison region carries implications for costs and timing. Hence most counterfactual approaches instead opt for the second option...
2. Comparison to an alternative scenario; essentially, trying to build a picture of what would have happened to the same product market in the same territory had the S&L programme not been introduced.

Whilst more practical, approaching the counterfactual in this way creates challenges which – the consensus from the available literature seems to be – are difficult to fully address:

- **Generating a baseline**: information on historic sales data from economies / sectors can vary considerably. If this data is poor, assumptions will likely need to be used and it becomes much more difficult to ensure an accurate sense of market status prior to the intervention.

- **Quantifying wider factors**: in the absence of a S&L programme, it is often assumed that the market for the affected product would have progressed towards greater energy efficiency; this might be due to various factors e.g. pre-existing S&L requirements or those in other countries that manufacturers are trading with[1], consumer demand / rising energy costs, wider government drives for sustainability, technological evolution and innovation, other pre-existing product policies or voluntary energy efficiency schemes. There is good awareness in studies of the range of factors that could be taken into account when building a counterfactual scenario - the challenge for evaluators is to disaggregate the likely effects of each of these and thereby arrive at an assessment of the specific influence of the programme. Time series data is used to address this issue (i.e. what was happening in the market prior to the S&L being introduced) but policy is rarely made in isolation of wider circumstances. Wilkenfeld (2008) highlights the example of a lighting MEPS whereby the market was *"already moving toward CFLs before the measure was announced and...CFLs already accounted for close to 50% of annual imports of GLS lamps."* He also cites several factors which could also be discerned at the time of the MEPS introduction, including a long-term trend towards CFLs from growing consumer familiarity, increasing quality and falling relative prices, as well as large-scale free giveaways and installations of CFLs in the period.

- **Assigning impact to individual elements of the same S&L programme:** linked to the challenge above, where MEPS and labelling are introduced concurrently, there is an additional challenge in separating the influence of each.

Looking beyond market performance

Whilst market and sales data are crucial to accurate evaluation of S&L programme impacts on energy and carbon savings, they are by no means the only data that should be considered within a robust evaluation.

- **Actual product performance:** a standard factor is often developed for a product within a certain efficiency band and this is used to assess the scale of energy efficiency improvement by sales of Product A being replaced by sales of Product B. However, these performance figures are sometimes obtained only through lab testing and it may be that actual field testing is needed to ensure this factor (and therefore extrapolated energy saving impacts) is accurate.

- **Compliance:** as highlighted in recent IEA policy pathways[2], MVE is an essential component of S&L programmes and evaluation should take account of the robustness of MVE and therefore the potential for non-compliance. In the absence of rigorous MVE, it may be that accuracy of labelling / claims of product performance / removal of non-compliant products is being exaggerated. A UK National Measurement Office study (2012) found that 22% of households were going on-line to purchase appliances, yet also found that less than half of products offered on-line had an accurate energy label. With on-line purchase likely to form an increasing proportion of sales, such a finding

[1] As noted in one study, *"while the proliferation of [S&L] programs is good for energy efficiency, it does make setting baselines somewhat of a guessing game."*
[2] www.**iea**.org/publications/freepublications/publication/monitoring-1.pdf

carries big implications for labelling programme impacts. Overall, assumptions of 100% compliance[3] will almost certainly lead to over-claim.

- **Consumer behaviour**: calculating energy saving impacts based upon the respective performances of the pre- and post-intervention product group makes the assumption that consumer behaviour (around purchasing and usage) has remained static. Yet, for example, consumers may be using a particular product for more time in a day than in previous years (or own more units of the product e.g. TVs), so whilst the 'per hour' usage has reduced, total usage has actually increased. Whilst this can be considered independent of the S&L programme in some cases, in others it may that steering consumers towards more efficient products has led to 'rebound' effects, which evaluators should take into account for accurate calculation of effects. It may be necessary for evaluators to analyse energy use in a sample of households both before and after the introduction of the programme. As an example of the unpredictability of consumer behaviour, Wilkenfeld (2008) highlights an instance whereby sales data indicated households had stockpiled the soon-to-be-removed inefficient types of light-bulb in the period leading up to the MEPS introduction.

- **Attribution to labelling**: MEPS removal of non-compliant products from the market leaves less room for ambiguity. Labelling programmes influence consumer choices over a longer period. For labelling programmes, an accurate assessment of the effect of the labels requires robust consumer surveying and a robust attribution methodology. Some consumers may have made the same purchase anyway due to other factors – e.g. price, innovative features, availability, appearance - and may not take any notice of the energy labelling. Even where the energy performance of products influences decisions, it may not have been the label itself which provided this information. Attribution factors should also take account of any retailer / manufacturer explanations as to how far they have actually attempted to promote or market products based upon the label.

Data quality and quantity

The quality of baseline and current market data is usually reliant upon the cooperation of manufacturers and / or trade associations. Reed and Hall () note that increasing market competitiveness can lead to industry reticence about providing full – or indeed any – data around product sales. Failure to obtain this creates risks arising from drawing conclusions based upon incomplete market data.

Linked to the challenge of obtaining data are risks around the quality of data. Some of this may be inadvertent – e.g. products and therefore their sales numbers not being assigned to the right energy usage band – but its effect upon calculations may still be significant.
Reed and Hall discuss the various challenges of obtaining the data required for proper assessment of market transformation and highlight that where programmes are regional rather than national, disaggregated data may not be available, whilst sales data *"for specific models of products are difficult to obtain from distributors and wholesalers."*

In addition, analysis of discussions with stakeholders as to the effects of S&L programmes upon market direction should always consider that respondents are not

[3] A further example noted in a recent report on the EU Energy Label by the UK body Consumer Focus: whilst the UK was highlighted in a Defra study as having *one of highest* compliance rates in the EU, three years after the label was created, compliance was still only found to be at 70%-80%.

disinterested. For example, hypothetically the effect of the programme may be downplayed by respondents where:

- They are nervous of policy success being used to justify further increases in the MEPS;
- They have a high performing product range and would benefit from implying that S&L requirements are not going far enough.

Implications for evaluation

Overall, a number of components are required to produce precise calculations of the impact of S&L programmes. The final section of this paper considers the implications of this for future evaluation of these programmes.

The types of challenges described above were discussed with evaluation colleagues; many of the issues and challenges were recognised and discussion generated some interesting considerations and further discussion points:

Do we need to go beyond current levels of robustness?

Whilst the potential for a more accurate calculation of directly attributable S&L programme impact is acknowledged, is such a change necessary?

It was noted in discussions that one of the key reasons S&L programme evaluation has not usually explored all the areas outlined above or sought to fully address all the limitations and challenges is that doing so would be largely irrelevant to the purposes of the evaluation.

Most S&L programmes are driven by the need to generate cost effective reductions in energy usage and carbon emissions and these programmes are deemed – and demonstrated to be - very cost effective. One experienced evaluator noted that whilst refinement of baseline / BAU scenarios might have a *"10% effect"* either way, the cost effectiveness performance of S&L programmes is so favourable that this would not affect policy maker decisions as to whether to introduce / continue the programme. Even accounting for a large margin of error, the sums still stack up.

Further to this, in constructing baseline / BAU scenarios, the literature indicates that evaluations are producing several different sets of assumptions which include a 'worst case' scenario, ensuring a very conservative, cautious assumption around the added value and impact of the S&L programme.

Discussions also highlighted that some potential issues are not as significant as they might appear to be in theory. For example, specific to the risk of misleading or non-existent manufacturer data responses, one experienced evaluator noted that the S&L programme teams and major players in the product market tend to work closely. This type of collaboration tends to ensure cooperation and minimise appetite for misleading responses; in fact it was pointed out that manufacturers are very aware of the risks of attempting this. An example was cited of industry in one country resisting the need for MEPS on the basis that the market was heading in the direction of increased efficiency anyway. The decision was therefore taken not to introduce a Standard; very soon afterwards, a low price, low efficiency imported product flooded the market and the resident industry lost market share.

Regarding disaggregation of impacts to different programmes, several sources in the literature state that this is very difficult and argue that it is valid to conclude that efficiency improvements have been caused by a combination of factors.

Can we go beyond current levels of robustness?

There is much to commend in current S&L programme evaluation; for example:

- There is usually consideration of multiple scenarios – ranging from best-case to cautious - when assessing the potential of a S&L programme e.g. NOMAD[4].
- In developing programme impact calculations, evaluations have generally used the most cautious savings estimate which assumes the highest level of market change that would have occurred regardless of the programme. Though it should be acknowledged that the status of this scenario as the most cautious is itself often reliant upon certain assumptions and limited data.
- Complex forecasting and impact assessment models are developed in an evidence-based way using actual market data. These often build in considerations such as the substantial improvement in product performance just prior to MEPS and labelling programmes being introduced.
- Some evaluations go beyond secondary data analysis and triangulate this with stakeholder and consumer surveys.

However, there are opportunities to enhance the robustness of S&L programme evaluation and impact assessment:

- Not all evaluations have explored and accounted for / addressed the gaps and risks highlighted in the sections above.
- All evaluations should acknowledge the challenges and limitations the approach used and report confidence intervals, standard deviations or, for qualitative data, more subjective rating systems.
- Although there are a growing number of studies seeking to assess the impacts of programmes already launched, the majority of S&L programme impact assessments are predictive and based upon scenario modelling. There would be value in conducting post-implementation primary and secondary research to test the assumptions and factors used in the original impact assessment in order to refine these.
- Secondary data provision can be variable and evaluations regarding S&L programmes in specific countries and sectors could be enhanced if more complete data were recorded and / or available.
- Many evaluations focus upon a programme specific to a product group or region, or explore only the 'standard' or the 'labelling' intervention; there would be value in consideration of multiple programmes within the same evaluation to address risks of double counting impact and provide more accurate disaggregation of observed energy consumption reductions / other impacts.

If we accept that there is the potential to go further in terms of robustness, there is inevitably the question of whether and how this would be enabled.

Stakeholders noted that some S&L programmes (especially where there is not substantial MVE) have small budgets, limiting the extent of evaluation activity (as already discussed, some programmes have yet to be evaluated following commencement).

Linking back to the need for more robust evaluation, another key barrier may be the appetite of the programme team (or their funders) for maximising the robustness of the evaluation.

First, in order to justify a place for the S&L initiative in legislative programmes, teams work through a lengthy process[5], often involving predictive modelling and stakeholder consultation. By the time the Standard or label has been successful in getting 'onto the

[4] Naturally Occurring Market Adoption (as referenced in Zhou, Romankiewicz, Vine, Khanna, Fridley 2012).
[5] *"Onerous and prescribed"* as one evaluator described.

books', teams may understandably feel that sufficient work has been done to test the costs and benefits.

Second, it is not clear from the programme team perspective that making the evaluation of S&L programmes more complex and in-depth is necessary. Stakeholders highlighted that most S&L programmes evaluation is adequate for what it is trying to do i.e. justify the existence of the programme. In most instances, the benefits outweigh the costs by such orders of magnitude that even substantial margins of error do not affect the overall case for funding / continuing the programme.

Future opportunities?

This review would seem to point to the view that even though there are ways to enhance S&L programme evaluation, there does not seem to be a need to fully address these. This is typical of all evaluation; sample sizes could always be bigger, attribution explored in more depth etc.

However, the energy efficiency policy landscape may precipitate a change in the way that S&L programme evaluation is conducted and viewed. One of the key considerations here is that whilst initial MEPS and labelling programmes for a product group may have had a profound effect, this impact *may* become more marginal; for example:

- Future increases in MEPS for a product group may produce more marginal benefits as the bottom 20-30% are much closer in efficiency performance to the market leading models than will have been the case for the original Standard.
- As consumers become increasingly concerned with energy costs, it may become more challenging to demonstrate the added value of some labelling (i.e. more consumers will feel that they were already going to specify / select the most efficient product anyway).

If potential benefits are becoming more marginal, this may necessitate closer consideration of some of the gaps and challenges described above, as the order of magnitude for programme impacts would be greatly reduced.

Yet this may not happen in the short term; it was noted in stakeholder discussions that if S&L programmes were not getting the support and endorsement they needed, there could be a requirement for more – and more robust – evaluation, but currently the reverse seems to be true: the number of programmes has steadily grown across the globe and existing ones are expanding (Wiel & McMahon 2001).

A further opportunity for more robust S&L programme evaluation may arise through emissions / carbon commodity trading, whereby more precise valuation of activity impacts is required.

Conclusions

- Evaluation of Standards & Labelling programmes carries a number of challenges and limitations, some of which are acknowledged in recent literature as very difficult to overcome, in particular the setting of precise baselines and considering the counterfactual scenario.

- A range of approaches have been designed and built upon to maximise the accuracy of S&L evaluation, though none appear to have *fully* addressed attribution of market change to the programmes being evaluated.

- However, current evaluation of S&L programmes can be assessed as strong on the basis that it delivers the data and level of robustness that governments / policy makers require. In addition, many approaches are carefully designed to generate as much accuracy as possible within the constraints of budgets and available data. For example, most evaluations attempt to build BAU scenarios and take into account a range of external factors in plotting this. In addition, where feasible, evaluations are looking to triangulate market sales data with stakeholder and consumer survey evidence.

- Overall, whilst there is a degree of uncertainty about the precise impacts and benefits of S&L programmes, there seems to be sufficient evidence on the *order of magnitude* of their effects to ensure their continuation and expansion.

- Some enhancements – e.g. increased post-launch testing and revisiting of assumptions, ensuring consumer and stakeholder responses are used to influence attribution – are possible but in the current policy climate are not required by most policy makers / programme teams and would be unlikely to alter appetite for these programmes.

- Yet there could develop circumstances in which the magnitude of impacts is reduced and more precise calculation of the benefits of S&L programmes (as well as the potential negative effects) is required.

References

Edward Vine 2008. *Strategies and policies for improving energy efficiency programs: Closing the loop between evaluation and implementation* -, Energy Policy, October 2008: 3872–3881.

Edward Vine, Peter du Pont, Paul Waide 2000. *Evaluating the impact of appliance efficiency labelling programs and standards: process, impact, and market transformation evaluations.*

Consumer Focus 2012. *Under the influence? Consumer attitudes to buying appliances and energy labels.*

Edith Molenbroek et al 2013. *First findings and recommendations: Evaluation of the Energy Labelling Directive and specific aspects of the Ecodesign Directive* - ENER/C3/2012-523

George Wilkenfeld 2008. *Evaluation of a lighting market transformation program in Australia – Outcomes and Attributions* - George Wilkenfeld and Associates, Sydney

Rod Ehler & Tom Talerico 1999. *It All Comes Down To the Baseline – Estimating Market Transformation Effects* – Quantum Consulting

Stephen Wiel & James E. McMahon 2001. *Energy-Efficiency Labels and Standards: A Guidebook for Appliances, Equipment and Lighting*

Nan Zhou et al 2012. *International Review of Frameworks for Impact Evaluation of Appliance Standards, Labeling, and Incentives* - China Energy Group Environmental Energy Technologies Division, Lawrence Berkeley National Laboratory - December 2012