Analysing controversies in energy policy: Assessing the evidence for rebound effects and global oil depletion

Submitted by Steve Sorrell to the University of Exeter as a thesis for the degree of Doctor of Philosophy by publication in Human Geography

March 2012

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: ..........................................................
Abstract

This thesis is submitted as a PhD by Publication. Part A provides an overview of the thesis and summarises its context, research questions, methodological approach and key findings. Part B is a collection of nine, first-named academic papers.

The thesis addresses two highly complex and controversial questions within energy policy, namely the nature and magnitude of ‘rebound effects’ from energy efficiency improvements and the extent and rate of depletion of global oil resources. Both of these questions are critically important to the development of a sustainable energy system and both are the subject of long-standing and highly polarised disputes. The thesis adapts, develops and applies a common methodology for reviewing the evidence on these questions, supplements this with original primary research and synthesises the results in a way that improves understanding and provides new insights.

The thesis includes four papers examining different aspects of rebound effects and four examining different aspects of global oil depletion. Given the complexity of the chosen topics, the papers cover a wide range of questions, issues and approaches. Collectively the papers: clarify relevant definitional and conceptual issues; evaluate competing methodological and analytical techniques; appraise the methodological quality of empirical studies; identify levels of uncertainty and potential sources of bias; develop simple mathematical models; conduct statistical analyses of primary data; compare and evaluate the results of modelling studies; and synthesise results from multiple research areas to provide novel insights into poorly understood phenomena. A ninth paper evaluates the strengths and limitations of systematic review techniques when applied to complex, policy-relevant questions such as these.

The thesis draws two main conclusions. First, rebound effects are frequently large and can substantially reduce the energy and carbon savings achieved from improved energy efficiency. Second, there is a significant risk that the global production of conventional oil will enter sustained decline before 2020. These conclusions run counter to conventional wisdom and have significant implications for public policy. The thesis also shows how the methodology of systematic reviews can be adapted and modified to make a valuable contribution to energy and climate policy research.
Acknowledgements

The research reported in this thesis was primarily funded by the UK Energy Research Centre (UKERC) which in turn is funded by the UK Research Councils. Additional funding was provided by the UK Economic and Social Research Council through their support for the Sussex Energy Group (SEG).

The research would not have been possible without the help of my colleagues in UKERC and SEG. I am particularly grateful to John Dimitropoulos for his invaluable contribution to the project on rebound effects, and to Jamie Speirs for his contribution to the project on oil depletion. Thanks also to Jim Skea, Rob Gross and Phil Heptonstall for their guidance, comments, patience and support.

The project on rebound effects benefited enormously from contributions by Grant Allan, Karen Turner, Matt Sommerville and the late Dennis Anderson, while the project on oil depletion benefited similarly from contributions by Adam Brandt, Richard Miller and Roger Bentley.

Many other colleagues contributed through sitting on advisory groups, commenting on early drafts, participating in discussions and providing data and information. These include Lester Hunt, David Broadstock, Peter McGregor, Kim Swales, Jake Chapman, Blake Alcott, Gordon Mackerron, Jim Watson, Horace Herring, Tina Dallman, Erica Thompson, Ken Chew, John Mitchell, Chris Skrebowski, Michael Smith, Paul Stevens, David Strahan, Jean Laherrère and Fabiana Gordon. I would particularly like to thank Harry Saunders for his insights and unceasing encouragement, and Manuel Frondel, and Robert Kaufmann for their careful peer reviews.

Many thanks to my supervisor Catherine Mitchell, for encouraging me to apply for a PhD by publication, for her supportive words and for insisting upon high standards!

Finally, special thanks to Angie for her love and support and to Rene and Layna for enriching my life. I dedicate this work to them.
Contents – Part A

INTRODUCTION ..................................................................................................................... 11

RESEARCH QUESTIONS......................................................................................................... 15

METHODOLOGY ..................................................................................................................... 15

CONTRIBUTION TO KNOWLEDGE ....................................................................................... 17
  CONTRIBUTION TO THE UNDERSTANDING OF REBOUND EFFECTS ................................ 17
  CONTRIBUTION TO THE UNDERSTANDING OF GLOBAL OIL DEPLETION ......................... 19

ORIGINALITY .......................................................................................................................... 20

INDEPENDENCE OF STUDY .................................................................................................... 22

ACADEMIC AND WIDER IMPACT .......................................................................................... 23

SPECIFIC CONTRIBUTION OF EACH PAPER .......................................................................... 25
  METHODOLOGICAL PAPER .................................................................................................. 25
    Paper 1 - Improving the evidence base for energy policy: the role of systematic reviews .... 25
  REBOUND EFFECT PAPERS ................................................................................................... 26
    Paper 2 - The rebound effect: Microeconomic definitions, limitations and extensions ......... 26
    Paper 3 - Empirical estimates of direct rebound effects: A review ....................................... 27
    Paper 4 - Energy-capital substitution and the rebound effect ............................................. 29
    Paper 5 - Jevons Paradox revisited: The evidence for backfire from improved energy efficiency ...... 30
  GLOBAL OIL DEPLETION PAPERS ....................................................................................... 31
    Paper 6 - Global oil depletion: A review of the evidence .................................................... 31
    Paper 7 - Oil futures: A comparison of global supply forecasts .......................................... 32
    Paper 8 - Hubbert's legacy: A review of curve-fitting methods to estimate ultimately recoverable resources ........................................................................................................ 33
    Paper 9 - Shaping the global oil peak: A review of the evidence on field sizes, reserve growth, decline rates and depletion rates .................................................................................................... 35

FURTHER RESEARCH ............................................................................................................. 36

REFERENCES ............................................................................................................................ 37
### Contents – Part B

| --- | --- |
Introduction

This thesis explores two major and long-standing controversies within energy policy - namely, the rebound effects from energy efficiency improvements and the depletion of global oil resources. It takes a common approach to these two controversies, based upon the methodology of systematic reviews [1].

Controversies pervade all fields of academic enquiry, but to varying degrees and with varying consequences for public policy. Controversies are especially prevalent when the relevant evidence is incomplete and uncertain, but they are also shaped by competing economic, political, and social interests that influence the questions asked, the weights attached to different types of evidence, the standards used to produce and evaluate that evidence, and the specific interpretations and judgements that are made [2]. Numerous studies have shown how values, interests and cultural practices can shape the production of knowledge, even in relatively ‘pure’ areas of scientific inquiry such as physics [3, 4]. But these factors play a much larger role in applied social science, where the research is commonly funded, produced and/or used by various interest groups; where there are difficulties in accessing relevant information and data; and where the subject is a complex and dynamic social system with all the attendant problems of interpretation and establishing causality [5]. These factors are especially relevant to energy and climate policy research, with the result that controversies in this field are common, vociferous and protracted.

Despite these difficulties, social scientific knowledge is relied upon by policy makers, who frequently need to make urgent decisions with far-reaching outcomes despite large uncertainties in the evidence base and conflicting recommendations by the relevant ‘experts’. Governments have responded to these difficulties in a variety of ways, but a particularly significant initiative has been the development of Evidence Base Policy and Practice (EBPP) - defined by Davies [6] as “…the integration of experience, judgement and expertise with the best available external evidence from systematic research”. This approach has its origins in the medical field, but was extended to public policy research by the former Labour government and has since been adopted by many other governments and international institutions [7-9]. In practice, EBPP has prioritised scientific research over other forms of evidence, given greater weight to quantitative and especially experimental research and focused in particular upon applying formal procedures to synthesise the evidence base – notably through the use of systematic reviews of the evidence [1]. In doing so, EBPP relies upon comparable and hence contestable judgements to those made within primary research and is therefore unlikely to provide a final resolution to any particular controversy. However, the formal procedures for synthesising research findings has improved understanding and increased consensus in many
areas of policy and practice, and the results of systematic reviews are frequently taken as the benchmark of current scientific understanding [10, 11]. This suggests that there should be merit in applying this approach to controversial questions within energy and climate policy, but for a variety of reasons it remains largely untried.

This thesis adapts the systematic review methodology and applies this approach to the topics of rebound effects and global oil depletion. It seeks to make a contribution to knowledge in both of these areas, as well as to the practice of systematic reviews.

The selected topics are critically important to the development of a secure and low carbon energy system. Oil depletion matters since oil accounts for a third of global primary energy supply and 95% of transport energy, with only limited scope for substitution in the short to medium term. An increasing number of commentators are forecasting an imminent peak and subsequent terminal decline in the global production of conventional oil, with non-conventional sources being unable to ‘fill the gap’ on the timescale required [13, 14]. Should this occur, it would severely damage the world economy for a decade or more [15], with greater impacts in importing countries since global export capacity is likely to decline more rapidly than global production [16]. To illustrate the magnitude of the challenge, most ‘peak oil’ forecasts anticipate that global production will fall by at least 2%/year - which is equivalent to the energy output of 100 nuclear power stations the size of Sizewell B (1 GW). But there is a long history of failed predictions of the ‘end of oil’ [17] and other commentators expect a combination of technical improvements in resource recovery, the development of non-conventional resources, improved vehicle efficiency and the diffusion of hybrid and electric cars to offset any potential supply constraints. The issues raised by this controversy have proved extremely difficult for governments to evaluate, with the result that policy responses to date have been muted and potentially inadequate

Improved energy efficiency is also critical for the development of a sustainable energy system. For example, in the ‘450 scenario’ of the IEA World Energy Outlook [18], improved energy efficiency accounts for 71% of the reductions in energy-related carbon emissions by 2020 and 48% by 2050. Not only does energy efficiency offer a large technical potential for reducing carbon emissions, it can also be highly cost effective and avoid the negative environmental impacts of many supply-side options. But improved energy efficiency can have multiple unintended consequences that have the potential to erode much of the anticipated energy

---

1 Conventional oil is defined in this thesis as including crude oil, condensate and natural gas liquids, but excluding oil sands and shales, coal to liquids, gas to liquids and biofuels [12].
savings. Analysts and policymakers tend to ignore these so-called ‘rebound effects’, but a growing body of academic research suggests they could be significant - with some analysts claiming that non-price measures to encourage energy efficiency provide an ineffective or even a counterproductive means of tackling climate change [21, 22]. As with oil depletion, the complex issues raised by this controversy have proved difficult for governments to evaluate, with the result that the policy responses to date have been equally limited.

The papers included in this thesis are listed in Table 1. The first paper evaluates the potential application of systematic reviews to energy policy questions, while the remainder adapt this approach to evaluate the evidence on each controversy and combine this with original primary research. Papers 2 to 5 evaluate the evidence on rebound effects and Papers 6 to 9 evaluate the evidence for global oil depletion. Eight of these papers are published in academic journals and one is a conference paper published online by the British Institute of Energy Economics. All of the papers were published or submitted between March 2007 and March 2012. During this period, I was a Senior Fellow in SPRU (Science and Technology Policy Research) at the University of Sussex and a participant in the UK Energy Research Centre (UKERC). All of the papers are based upon my earlier publications for UKERC [12, 23-32] and have been selected from a longer list of publications that represent my academic work over this five-year period.

---

2 For example, they are not mentioned by the Stern Review of the Economics of Climate Change [19] or by the Intergovernmental Panel on Climate Change [20].
**Table 1 Papers included in the thesis**

<table>
<thead>
<tr>
<th>Paper</th>
<th>Authors</th>
<th>Reference</th>
</tr>
</thead>
</table>
**Research questions**

The research questions addressed by this thesis are as follows:

- **Systematic reviews**: What contribution can systematic reviews make to energy policy research and how can these techniques be adapted to enhance their contribution?
- **The rebound effect**: What is the evidence that improvements in energy efficiency will lead to economy-wide reductions in energy consumption?
- **Global oil depletion**: What evidence is there to support the proposition that the global supply of conventional oil will be constrained by physical depletion before 2030?

The second and third research questions have much in common. They concern complex and multi-faceted issues whose evaluation involves the integration of multiple sources of evidence on a variety of topics and the development of knowledge in a number of areas. They are the subject of long-standing, contentious, and highly polarised disputes that reflect in part disciplinary boundaries (e.g. economists versus geologists). They are characterised by confusion and disagreement over key concepts, definitions and causal mechanisms, together with related disputes over the relevance and credibility of different analytical techniques. They suffer from serious gaps and limitations in the available data and evidence base. And they concern politicised and increasingly high profile issues that are critical to future economic welfare and environmental sustainability. Not all questions relevant to energy policy have these features, but a very large number of them do. Hence, the thesis examines how the systematic review methodology can improve understanding and increase the degree of consensus on complex and controversial questions such as these.

**Methodology**

The research methodology was informed by the well-established practice of systematic reviews, but modified this in a number of ways. Systematic reviews use an explicit, transparent and replicable methodology to identify, appraise, select and synthesize all high quality research evidence relevant to a particular research question [33]. Core features of this approach - which is described and evaluated in Paper 1 - include precise specification of the review question, exhaustive searching of the available literature, careful appraisal of the methodological quality of different studies and reliance upon the more rigorous studies when drawing conclusions. The thesis adapts this approach to make it more suitable for the complex and multidimensional questions that are of interest within energy policy, together with the types of evidence that is commonly available. For example, systematic reviews rely heavily
upon (quasi-) experimental studies, but these are relatively rare in the energy field. Instead, the evidence base is dominated by econometric analyses of secondary data and energy-economic modelling.

A common approach was taken to both topics. The first stage was to prepare a Scoping Note that identified the key features of the academic and policy debate, the major sources of controversy, the size, nature and characteristics of the evidence base and the potential contribution of an evidence-based assessment [34, 35]. This was circulated to stakeholders for comment and a 5-8 member Advisory Group was formed, representing a diversity of perspectives and expertise on the relevant issues. The scope of the assessment was then refined in consultation with this Advisory Group. For example, in the case of oil depletion, I chose to focus upon the physical and geological factors governing the size of the recoverable resource of conventional oil and the rate at which it can be brought into production in the period up to 2030.

The second stage was to prepare and publish an Assessment Protocol that specified in detail the objectives of the assessment, the specific sub-questions that would be addressed, the sources of evidence relevant to those questions and the criteria by which these would be evaluated. For example, in the case of oil depletion I identified 27 sub-questions covering issues such as the definition and interpretation of reserve estimates, the nature and importance of reserve growth and the strengths and weaknesses of different modelling tools.

The third stage was to systematically search the academic and grey literature on the relevant topics, using multiple keyword searches of several databases. The results were then filtered according to methodological and other criteria and classified according to their relevance to the identified sub-questions.

The fourth stage was to critically evaluate the evidence on each of these sub-questions, focusing in particular upon the methodological robustness of each study and placing greater weight upon the more rigorous studies. With approximately 500 sources of evidence for each topic, this was a substantial task. Given the nature and diversity of this evidence, the standard systematic review procedures for assessing methodological quality could not be applied. Instead, the primary focus was upon clarifying multiple definitional and conceptual issues and highlighting the strengths and weaknesses of different techniques. In addition, several questions were investigated through the analysis of primary data sources - most notably a confidential database of regional oil exploration, discovery and production provided by IHS Energy [36].
The fifth stage was to draft a number of technical reports on the relevant issues and to circulate these to the Advisory Group and others for comment [24-31, 37, 38]. The final stage was to integrate these findings into an overall synthesis report [12, 23], have this peer reviewed and to publish and disseminate the results.

**Contribution to knowledge**

The evaluation of the evidence for both topics involved a wide range of activities, including: clarifying relevant definitional and conceptual issues (e.g. Papers 2, 3 and 8); evaluating competing methodological and analytical techniques (e.g. Paper 8); appraising the methodological quality of empirical studies (e.g. Paper 4); identifying levels of uncertainty and potential sources of bias (e.g. Papers 2 and 8); and synthesising results from multiple research areas, with the aim of providing new insights into poorly understood phenomena (e.g. Paper 5). This process went well beyond the straightforward comparison and meta-analysis of quantitative results that are normal for a systematic review. I consider that this process of conceptual development and creative synthesis provides a contribution to knowledge in each of the subject areas.

The assessment process also extended to several activities more commonly classified as primary research. The aim was to improve understanding of particularly contentious and confused issues and fill identified ‘gaps’ in the current state of knowledge. These activities included the development of simple mathematical models (Papers 2 and 3), the statistical analysis of industry data on regional oil production and discovery (Papers 8 and 9) and the systematic comparison and evaluation of global oil supply forecasts (Paper 7). I consider that the results of this research provide an additional contribution to knowledge.

In addition, the thesis provides new and valuable insights into the strengths and limitations of systematic review techniques when applied to complex, policy-relevant questions such as these.

The following two sections summarise the contribution to knowledge in each of the topic areas.

**Contribution to the understanding of rebound effects**

The potential ‘energy savings’ from improved energy efficiency are commonly estimated using engineering-economic models, but these neglect the impact of important behavioural responses to such improvements, such as increased demand for cheaper energy services. These responses typically act to reduce the energy savings achieved. A variety of mechanisms contribute to such ‘rebound effects’, but their net effect is difficult to quantify and is widely
ignored. The research sought to improve understanding of the nature and size of these rebound effects and to examine whether they could be sufficiently large as to increase energy consumption at the level of the national or global economy (‘backfire’). Papers 2-5 report the main findings (see also [23]).

The research showed how quasi-experimental and econometric techniques have yielded relatively robust estimates of ‘direct’ rebound effects following improvements in the energy efficiency of household energy services in the OECD. These effects result from increased demand for the relevant energy services (e.g. heating, travel) and can erode 10-30% of the potential energy savings before other indirect and economy-wide effects are accounted for. Direct rebound effects are likely to be greater in developing economies and also appear more significant in industry, although here the evidence is much weaker.

Quantification of the various indirect and economy-wide rebound effects is more difficult, but insights may be gained from energy-economic models of the macro-economy. The available studies relate solely to energy efficiency improvements by producers and show that the economy-wide rebound effect varies widely depending upon the nature of the energy efficiency improvement and the sector in which it takes place and is sensitive to wide range of variables. All studies conducted to date estimate economy-wide effects in excess of 30% and several predict backfire. Moreover, these estimates do not take into account the amplifying effect of any associated improvements in the productivity of capital, labour or materials.

Quantification of rebound effects is hampered by inadequate data, unclear system boundaries, endogenous variables, time-delayed feedback loops uncertain causal relationships, trans-boundary effects and complex, long-term dynamics such as changing patterns of consumption [39]. Since economy-wide effects are emergent phenomena resulting from the complex interaction of multiple actors and mechanisms, studies of a subset of mechanisms within narrow spatial and temporal boundaries can provide only a partial picture. Generally, as the boundary and scope of analysis expands the estimated size of rebound effects increase. The risk of backfire appears especially high for energy efficiency improvements associated with the early stage of diffusion of ‘general-purpose technologies’ such as engines, motors and computing. These have historically accounted for a significant proportion of total energy consumption.

The research highlighted both theory and evidence suggesting that: a) there is relatively limited scope for substituting other inputs for energy at the aggregate level; b) that technical change has frequently acted to increase energy intensity; c) that weighting energy carriers by their relative economic productivity leads to different conclusions regarding the extent to
which energy consumption has been decoupled from economic output (and the reasons for that decoupling); and d) that increases in the quantity and quality of energy inputs, together with the efficiency with which they are used, can have synergistic and multiplicative impacts on the productivity of capital and labour. All these observations point to energy playing a more important role in economic growth than is conventionally assumed and may be used in support of arguments that rebound effects are large. But the research identified numerous flaws in these arguments, together with gaps and limitations in the empirical evidence used in their support. Hence, backfire has not been convincingly demonstrated to be a universal outcome of cost-effective energy efficiency improvements.

The research demonstrates that the potential contribution of energy efficiency improvements to emission reduction needs to be reappraised and that rebound effects need to be taken into account when developing and targeting energy efficiency policy. Rebound effects may frequently be large, but can be mitigated through carbon/energy pricing and the imposition of carbon caps.

**Contribution to the understanding of global oil depletion**

Conventional oil resources are substantially depleted and many commentators forecast a near-term peak and subsequent terminal decline in global production. More optimistic observers note how previous forecasts of a global peak in production have proved incorrect and emphasise the potential for improved resource recovery and fuel substitution. This polarised debate has inhibited a balanced appraisal of the issues involved. Focusing primarily upon ‘physical’ factors, the research sought to improve understanding of the mechanisms of oil depletion, to identify the extent and pace of depletion at the global level and to assess the relevant risks and uncertainties. Papers 6-9 report the main findings (see also [12]).

The research showed that the mechanisms leading to a ‘peaking’ of conventional oil production are well understood and provide identifiable constraints on future supply at both the regional and global level. Fundamental features of the conventional oil resource make it inevitable that production in a region will rise to a peak or plateau and ultimately decline. These include the production profile of individual fields, the concentration of resources in a small number of large fields and the tendency to discover and produce these fields relatively early. This process can be modelled and the peaking of conventional oil production can be observed in an increasing number of regions around the world.

Public domain data sources are poorly suited to studying depletion and frequently misinterpreted while industry data is confidential and not necessarily reliable. But despite large uncertainties, sufficient information is available to allow the status and risk of global oil
depletion to be adequately assessed. In particular, there is scope for greater consensus on long-standing controversies such as the source and magnitude of ‘reserve growth’ at existing fields.\(^3\) This currently makes a larger contribution to global reserve additions than the discovery of new fields, but is likely to decline in importance in the future. The key challenge is the concentration of global production in a small number of ageing ‘giant’ fields, many of which are past their peak of production. Simply to prevent global production from falling requires investment in new production capacity equivalent to a new Saudi Arabia every 3-4 years. With accelerating decline from currently producing fields, falling discovery rates and greater reliance upon smaller, deeper and more complex fields in challenging locations, it is becoming increasingly costly and difficult to maintain global supply.

Methods for estimating resource size and forecasting future supply have important limitations and the corresponding uncertainties in such estimates are insufficiently acknowledged. Increasing the complexity of the relevant models does little to reduce these uncertainties and is subject to rapidly diminishing returns. Commonly used ‘curve-fitting’ methods are unreliable and likely to underestimate recoverable resources and provide excessively pessimistic forecasts of future supply. However, while the global recoverable resources of conventional oil are likely to be greater than many estimates suggest, these will be slow, expensive and difficult to produce even when political and economic conditions prove favourable and may make little difference to the timing of the global peak.

The research suggests that conventional oil production is likely to enter a sustained decline before 2030 and there is a significant risk of this occurring before 2020. Given the lead times required to both develop substitute fuels and improve energy efficiency, this risk deserves urgent consideration

**Originality**

The process of assessing the evidence on these topics was partly akin to assembling a jigsaw in which multiple pieces were missing – with the synthesis providing an overall picture and the supporting primary research filling in some gaps. However, no one ‘true’ picture emerged. Not only was the evidence base patchy, multi-faceted and ambiguous, but the identification and interpretation of this evidence hinged upon contested theoretical assumptions - such as the suitability of neoclassical production functions for modelling producer behaviour, or the relative weight to give to physical and economic variables when projecting future oil supply.

\(^3\) The growth in the estimates of ultimately recoverable resources from a field or region, as a consequence of better geological understanding, improvements in extraction technology, variations in economic conditions, changes in reporting practices and other factors.
These characteristics precluded any straightforward judgement on the research questions and led instead to an emphasis upon clarifying the issues and mechanisms involved, exposing underlying theoretical assumptions and establishing the reasons for the competing interpretations. Nevertheless, it was possible to form judgements on the credibility of specific theoretical assumptions and the degree of support for particular empirical claims. It was also possible to highlight interpretations that appeared at variance with the evidence base, to identify areas where there was greater scope for consensus and to prioritise areas for further research. I consider that this process of synthesis, together with the supporting empirical research, have provided an original contribution to knowledge. I illustrate this with the help of five examples from the thesis.

- **Defining and estimating the direct rebound effect**: The direct rebound effect is relatively well researched, but the existing studies are hard to compare and employ competing measures and definitions. Paper 2 develops a unified theoretical framework that clarifies the relationship between these definitions, as well as identifying the nature and consequences of various sources of bias. Both can be used to interpret existing and guide future empirical research. The paper also provides a novel analysis of the relevance of the opportunity cost of time to rebound effects and highlights the existence and potential importance of rebound effects from time-saving innovations.

- **Clarifying energy-capital substitution**: The economic literature on the scope for substitution between energy and capital is difficult to interpret owing to multiple definitions, problematic assumptions and inappropriate use of empirical results. Paper 4 clarifies this literature and uses this to contest the claim that rebound effects will be larger when substitution between energy and capital is easier [40]. It shows how an empirical finding that energy are ‘complements’ at a particular level of aggregation [41, 42] may be consistent with rebound effects that exceed 100% (backfire).

- **Linking Jevons Paradox and ecological economics**: The claim that cost-effective energy efficiency improvements will lead to an increase in aggregate energy consumption (‘Jevons Paradox’) is difficult to investigate and poorly understood. Paper 5 illustrates the multiple dimensions of this issue and shows the importance of neglected factors such as the ‘quality’ of energy carriers and the links between energy productivity and total factor productivity. In doing so, it introduces a range of issues not previously discussed in this context and makes some novel connections between arguments in favour of Jevons Paradox and heterodox claims regarding the contribution of energy to productivity improvements and economic growth.

---

4 For example, holding output constant the demand for capital falls when energy prices rise.
Comparing global oil supply forecasts: The ‘peak oil’ debate hinges upon competing forecasts of future supply, but these have not been systematically compared. Paper 7 provides such a comparison and shows how the different forecasts rely upon different implicit or explicit assumptions for a limited range of key variables, including the ultimately recoverable resources of conventional oil. By comparing these assumptions against the available evidence on these variables, the paper is able to assess the relative plausibility of different forecasts and the consequent risk of a near-term peak in global production.

Testing the reliability of curve-fitting methods: Fossil fuel resource estimates and forecasts of future supply are frequently derived by fitting curves to time-series data on regional production or discovery and projecting these forward in time. Paper 8 uses illustrative data from seven oil-producing regions to demonstrate the unreliability of these techniques. The papers show how widely different results may be obtained from different techniques, functional forms, lengths of data series and numbers of curves [for details, see 29]. For example, two functional forms are shown to fit the discovery data for one region equally well (a difference of only 0.003 in $R^2$) but the resulting resource estimates differ by a third. These papers provide the first systematic demonstration of the limitations of such techniques.

Independence of study
I led the two UKERC projects on which this thesis is based and was responsible for the design and conceptual development of each, the planning and execution of the detailed work programmes and the delivery of reports. I was the sole author of the synthesis report on the rebound effect [23], the primary author of the synthesis report on global oil depletion [12], primary author of three technical reports that provided the foundation of each project [24, 25, 29] and co-author of seven of the other nine technical reports [26-32]. I also organised the contributions from the Advisory Groups, led the communication with experts and peer reviewers and disseminated results to academics, policy makers and other audiences.

I was ably assisted by John Dimitropoulos in the assessment of rebound effects and by Jamie Speirs in the assessment of global oil depletion. Both were responsible for the initial literature search and for a variety of research, administrative and drafting inputs throughout the assessment process. This included John’s assistance with the analysis and evaluation of econometric studies of rebound effects (Papers 2 and 4) and Jamie’s assistance with the statistical analysis of oil production and discovery data (Paper 8). I also commissioned two contributions from external researchers [37, 38]. But in all areas, I provided the main intellectual contribution, was responsible for reviewing and synthesising the bulk of the
evidence and was the primary or sole author of the subsequent outputs. As a result, I am the sole or lead author on all the included papers.

**Academic and wider impact**

An indication of the academic impact of this thesis is provided by Table 2 which shows the cumulative number of citations to the eight papers published in academic journals over the period March 2007 to January 2012, together with the mean number of citations per year. Paper 2 is a conference paper and hence excluded from this table, while Paper 9 has yet to be published. In total, the papers included in this thesis received 270 citations in Google Scholar over this period, 103 in Scopus and 67 in ISI Web of Science, with each paper being cited an average of 13 times a year in Google Scholar.\(^5\) However, the mean period since the publication of these papers and the submission of this thesis is only 27 months, with Papers 6-8 being published in September 2010 and Paper 9 being published in January 2012.\(^6\)

A second indication of academic impact is provided by the annual *impact factor*, defined as the average number of citations received to papers published during the preceding two years. Using ISI, the included papers had an average impact factor of 4.5 in 2009, 4.0 in 2010 and 3.1 in 2011, which compares to a mean JCR impact factor of 2.8 for the most prominent journals in this area.\(^7\) Using Scopus, the included papers had an average impact factor of 7.0 in 2009, 5.5 in 2010 and 7.9 in 2011.

Most of these citations are to the three journal papers on rebound effects (Papers 2, 4 and 5), and in particular to Paper 2 which received 98 citations in Google Scholar and 36 in Scopus over this period. The conceptual framework introduced by Paper 2 forms the basis for an increasing amount of empirical research in this area [e.g. 44, 45-48]. Papers 6-9 on oil depletion have received fewer citations to date, but these were only published in September 2010. My work on oil depletion has led to an invitation from the Royal Society to edit a special edition of the *Philosophical Transactions A* on the future of global oil supply. The impact of the research on non-academic audiences is summarised in Box 1.

---

\(^5\) Or 15 times a year if Paper 9 is omitted.
\(^6\) In addition, the synthesis report on the rebound effect [23] has received 144 citations on Google Scholar, the synthesis report on global oil depletion [12] has received 40 citations and my co-edited book on the rebound effect [43] (to which I contributed four chapters) has received 38.
\(^7\) *Energy Policy* (3.02), *Energy* (3.65), *Energy Journal* (1.85) and *Energy Economics* (2.9)
Table 2 Citations to papers included in this thesis and published in academic journals over the period March 2007 to January 2012

<table>
<thead>
<tr>
<th>Paper</th>
<th>Date of publication</th>
<th>Cumulative no. of citations since publication</th>
<th>Mean no. of citations per year since publication (Google Scholar)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Google Scholar</td>
<td>Scopus</td>
</tr>
<tr>
<td>Paper 1</td>
<td>March 2007</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>Paper 2</td>
<td>April 2008</td>
<td>98(^b)</td>
<td>36</td>
</tr>
<tr>
<td>Paper 4</td>
<td>April 2009</td>
<td>80</td>
<td>32</td>
</tr>
<tr>
<td>Paper 5</td>
<td>April 2009</td>
<td>42</td>
<td>14</td>
</tr>
<tr>
<td>Paper 6</td>
<td>September 2010</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>Paper 7</td>
<td>September 2010</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>Paper 8</td>
<td>September 2010</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Paper 9</td>
<td>January 2012</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>270</td>
<td>103</td>
</tr>
</tbody>
</table>

Notes:

a. Includes self-citations

b. Excludes 25 citations to an earlier version of this paper presented at the 29th IAEE international conference in Potsdam, June 2006
Box 1 Wider impacts of the research

The results of the rebound effect project [23] were published in October 2007 and attracted widespread media attention, including BBC radio, four UK national newspapers, several UK local radio stations, the Economist magazine and a large number of websites and trade publications. Media interest has continued, including coverage in the Guardian, New York Times [49], New Yorker [50], Conservation magazine [51], Scientific American and Nature Climate Change [52]. Follow-up activities have included talks and conference presentations in the UK, US, Canada, Germany, Belgium and Austria, together with meetings with the UK Department of Energy and Climate Change (DECC) and the Department of Environment, Food and Rural Affairs (DEFRA). The results from the study have been included in policy guidance from the UK government [53], and the synthesis report has formed the foundation for follow-up studies by the European Commission [54] and the US Breakthrough Institute [22].

The results of the global oil depletion project [12] were published in October 2009 and attracted comparable media attention, including BBC TV and radio, Reuters, six UK national newspapers, articles in Science [55], The Ecologist, New Scientist, Power UK and other publications and extensive online coverage. The project results have since been presented at numerous conferences and meetings, including three with DECC, one with the Irish Sustainable Energy Authority (SEAI) and an invited submission to a DECC inquiry, led by the Chief Scientist, David Mackay. The study has also been cited in technical publications such as IMF World Economic Outlook 2011.

Specific contribution of each paper

This section summarises the main findings and contribution of each paper. The papers are organised into three groups, namely: methodology, rebound effects and global oil depletion.

Methodological paper

Paper 1 - Improving the evidence base for energy policy: the role of systematic reviews

This paper introduces the concept of systematic reviews to an energy policy audience and evaluates their potential contribution to research and practice in this area. The paper identifies the rationale for and concept of a systematic review, summarises the methods through which it is commonly achieved and discusses its advantages and limitations. It then examines the particular challenges of applying this methodology to energy policy questions.

The paper argues that energy policy research has many of the problems that a systematic reviews set out to address, such as conflict and confusion over key issues, over-reliance on individual studies, inadequate accumulation and synthesis of results, insufficient attention to methodological quality, conflicting recommendations by different authors and uncertainty...
over whom to trust. Systematic reviews set out to address these problems, but they have several drawbacks of their own, including the narrow range of questions to which they have been applied, the bias towards quantitative (and especially experimental/quasi-experimental) research methodologies, the difficulties in addressing complex problems and policies, and the ‘additive’ approach to synthesis that neglects the complementary nature of different studies and perspectives. The paper identifies the implicit philosophy of science as the source of these limitations and shows how competing philosophies of science (notably critical realism [56]) would lead to differing emphases and alternative approaches.

The paper then identifies a number of important differences between the types of question and evidence for which systematic reviews have been most successful and the type of question and evidence that are most common within energy policy. For example, systematic reviews rely primarily upon ex-post evaluations and rarely use ex ante modelling studies, but the former are rare within energy policy research (despite their potential in areas such as energy efficiency policy) while the latter are dominant. The implications of these differences is illustrated through the example of rebound effects, where the paper concludes that the ‘narrow’ systematic review methodology is only appropriate for assessing a subset of the relevant evidence, and even then is difficult to apply. As a consequence, the paper suggests that energy researchers will only be able to apply the traditional systematic review methodology to a limited range of questions and will probably need to modify and extend the approach when it is applied.

This paper informed the approach taken to the two assessments and subsequent experience has reinforced the paper’s conclusions. While some features of systematic reviews were retained (e.g. systematic searching of the relevant literature; evaluating the methodological quality of studies), the overall approach was much more flexible and adaptable.

Rebound effect papers

Paper 2 - The rebound effect: Microeconomic definitions, limitations and extensions
The focus of this paper is the definition of direct rebound effects and the estimation of such effects through the econometric analysis of secondary data. Studies estimating such effects are difficult to compare owing to competing definitions, inconsistent notation and varying populations, measures, datasets, methodologies and controls. To bring some clarity to this area, the paper sought to provide a rigorous definition of the direct rebound effect, clarify key conceptual issues and highlight the potential consequences of various assumptions for the accuracy of econometric estimates.
The paper uses Becker’s ‘household production’ framework [57] to describe consumers’ demand for energy services and to illustrate the trade-offs involved. It shows how the direct rebound effect can be represented as the elasticity of energy demand with respect to energy efficiency and how this may be decomposed into the sum of elasticities for the number, capacity and utilisation of energy conversion devices. The paper then explores the relationship between this measure and those more commonly used to estimate the direct rebound effect, namely: a) the elasticity of demand for energy services with respect to the energy cost of energy services; b) the elasticity of demand for energy services with respect to the price of energy; and c) the elasticity of demand for energy with respect the price of energy. It identifies the assumptions required for these measures to be equivalent, their expected relative magnitude, the advantages and disadvantages of each as a measure of rebound effects and their application in practice.

The paper then exposes the limitations of three underlying assumptions of the ‘price-based’ definitions. First, it argues that capital costs form an important part of the total cost of providing energy services and shows how empirical studies that neglect this are prone to bias. Second, it shows why energy efficiency should be treated as an endogenous variable, derives an alternative definition of the direct rebound effect that allows for endogeneity and argues that empirical studies, would frequently benefit from using simultaneous equation models. Third, it explores the implications of the opportunity costs of time in the production of energy services and highlights the consequences for energy use of improved time efficiency, the influence of time costs on direct rebound effects and the existence of a parallel rebound effect with respect to time. Each of these considerations serves to highlight the difficulties in obtaining reliable estimates of direct rebound effects and the different factors that need to be controlled for. Finally, the paper highlights the implications of these findings for econometric studies and argues that many existing studies are likely to overestimate such effects.

The value of the conceptual synthesis provided by this paper is demonstrated by the large number of subsequent citations (Table 2) and the further development of the ideas by Greene [44] and Frondel et al. [46, 58].

**Paper 3 - Empirical estimates of direct rebound effects: A review**

This paper complements Paper 2 by evaluating existing estimates of the direct rebound effect for household energy services. The estimates are derived from both quasi-experimental studies and the econometric analysis of secondary data. The paper pays careful attention to how the direct rebound effect is defined and measured, the quality of the data used, the
robustness of the methodologies employed, the potential sources of bias and error and the applicability of the results to other populations and time periods.

The paper shows that the estimation of direct rebound effects is far from straightforward and the required data is frequently lacking. As a result, the evidence is sparse and methodologically diverse and largely confined a small number of household energy services in OECD countries. The methodological quality of many quasi-experimental studies is poor, while many of the econometric estimates may be upwardly biased. Since the own-price elasticity of energy demand provides an upper bound, the direct rebound effect for most household energy services in the OECD should be less than 100%.

Using the more rigorous studies, the paper concludes that the mean value of the long-run direct rebound effect for personal automotive transport is less than 30% and likely to be closer to 10%. Available estimates are broadly consistent, despite wide differences in populations, datasets, methodologies and elasticity measures, suggesting that the results are robust. The evidence on household heating and cooling is more diverse and less reliable, but points to a mean value of around 20% with significantly higher values for low income groups. Comparison of estimates is hampered by varying choices of dependent and independent variable and confusion between ‘shortfall’, ‘temperature takeback’ and ‘behavioural change’. Both theory and limited empirical evidence suggest that rebound effects are smaller for other consumer energy services and are likely to decline in the future as demand saturates and incomes increase.

The paper provides several important qualifications to these conclusions, including the relatively limited time periods over which these effects have been studied and the restrictive definitions of energy services that have been employed. For example, current studies only measure the increase in distance driven for personal automotive transport and do not capture any efficiency-induced increases in vehicle size, weight and power. They also fail to capture the economic and social transformations that cheaper energy services may induce over the very long term [59]. Rebound effects for space heating and other energy services are also higher among low-income groups and most studies do not account for ‘marginal consumers’ acquiring energy services such as space cooling for the first time. This suggests that direct rebound effects for household energy services could be substantially higher in developing countries.

---

8 ‘Shortfall’ refers to the gap between anticipated and measured energy savings; ‘temperature take-back’ refers to the increase in internal temperatures following efficiency improvements; and ‘behavioural change’ refers to behavioural responses by building occupants, such as increased thermostat settings. These measures are not equivalent.
The evidence base has grown significantly since this paper was published, reflecting growing interest in the topic [48, 60-65]. With the notable exception of Frondel et al. [45, 58], these more recent studies have produced estimates that are consistent with the above conclusions.

**Paper 4 - Energy-capital substitution and the rebound effect**

The focus of this paper is the scope for substituting between energy and capital in economic production and the relevance of this to rebound effects. Both theoretical and empirical work by Saunders [40, 66] and a growing number of CGE modelling studies [67-70] suggest that rebound effects are larger when substitution between inputs is easier. But while energy and labour are generally found to be substitutes, more than one hundred studies over a period of three decades have failed to reach a conclusion on whether energy and capital may be regarded as ‘substitutes’ or ‘complements’ [71]. This paper attempts to make sense of this literature, identifies reasons for the diverging conclusions, questions the usefulness of the relevant concepts and argues that the relationship between substitution and rebound effects is more complex than commonly assumed.

The paper begins by clarifying the meaning of substitution, summarising the multiple definitions of the elasticity of substitution and arguing that the definition most commonly used is of little practical value. It shows the importance of the level of aggregation at which substitution is measured and why the standard assumption of ‘separable’ inputs can bias results. The paper then shows why there is only a tenuous link between empirical estimates of substitution elasticities and the assumptions used by CGE models - owing to the latter using different production functions, separability assumptions and elasticity definitions from empirical studies and employing estimates from inappropriate sectors, time periods and/or levels of aggregation.

The paper then explores the relationship between elasticities of substitution and the rebound effect. It shows how this relationship hinges upon the distinction between energy and energy services, the choice of production function, the definition of the elasticity of substitution and the validity of the assumption of separability between energy services and other inputs. When these are taken into account, it is possible for rebound effects to be large for a given efficiency improvement, sector and time period even when empirical studies find energy and capital to be complements. This paper demonstrates this by reinterpreting the results of a seminal study by Berndt and Wood [41] and showing how this can be understood as an illustration of backfire following an energy efficiency improvement, despite their finding of energy-capital complementarity.
This paper was presented at the BIEE 2008 academic conference, but was rejected by *Energy Policy* for being too technical and by *Energy Economics* for lacking focus - despite one review being extremely positive.

**Paper 5 - Jevons Paradox revisited: The evidence for backfire from improved energy efficiency**

This paper evaluates the claim that cost-effective energy efficiency improvements will increase rather than reduce economy-wide energy consumption (‘Jevons Paradox’). This proposition was first advanced by William Stanley Jevons in 1865, but is very difficult to investigate empirically. While substantial improvements in energy efficiency typically occur alongside increases in economic output, total factor productivity and overall energy consumption, the causal links between these variables are difficult to isolate - in part because each contributes to other within positive feedback loops [72].

Proponents of Jevons Paradox support their case through a mix of theoretical argument, illustrative examples and ‘suggestive’ evidence from econometric analysis and economic history. The paper reviews these arguments and evidence, together with a number of others that appear relevant including the definition and measurement of energy and energy productivity, applications of neoclassical production and growth theory, econometric analysis of productivity improvements and the direction of technical change and estimates of the embodied energy associated with efficiency improvements. It argues that disputes over the size of the economy-wide rebound effect can be linked in part to competing views on the contribution of energy to productivity improvements and economic growth.

The paper concludes that the theoretical arguments used by key proponents of Jevons Paradox rely upon stylised models with numerous limitations, while the empirical evidence cited in support of the universal applicability of the Paradox is both selective and ambiguous. At the same time, these arguments and evidence deserve more attention than they have received to date. The paper shows why rebound effects vary with the nature and location of the energy efficiency improvement and how they represent the outcome of multiple and reinforcing mechanisms that are likely to increase the magnitude of such effects over space and time. ‘Win-win’ energy efficiency improvements that simultaneously reduce non-energy costs and/or improve the productivity of other inputs are likely to have large rebound effects and appear to be common.

The paper highlights the synergies between several of the arguments in support of Jevons Paradox and the heterodox claim – derived from the ‘biophysical school’ of ecological
economics - that the increased availability and quality of energy inputs is the primary driver of total factor productivity improvements and economic growth [73, 74]. It summarises some arguments and evidence in support of this proposition, but also highlights several flaws. The paper argues that the debate over Jevons’ Paradox would benefit from more careful distinctions between different measures and types of energy-efficiency improvement and greater attention to the importance of energy quality. Jevons’ Paradox is more likely to hold for energy-efficiency improvements associated with the early stage of diffusion of ‘general-purpose technologies’ (such as electric motors, lighting and computers) which stimulate new energy-using applications, products and industries and have such significant effects on innovation, productivity and economic growth that economy-wide energy consumption is increased. But it is less likely to hold for the later stages of diffusion of these technologies, or for ‘dedicated’ energy-efficiency technologies such as improved thermal insulation. While understanding of these issues remains limited, the paper identifies several promising avenues for research and argues that the neglect of rebound effects by researchers and policymakers can no longer be justified.

The debate on Jevons Paradox has grown in prominence since this paper was published as well as becoming more nuanced. This assessment appears to have influenced the debate since the relevant articles reference either the synthesis report or this paper [e.g. 50, 51, 75].

**Global oil depletion papers**

**Paper 6 - Global oil depletion: A review of the evidence**

This paper summarises the main conclusions of the assessment of global oil depletion. It begins by clarifying some relevant definitions, then highlights the uncertainty and limitations of the available data sources, explains the mechanisms contributing to the ‘peaking’ of regional oil supply and summarises global trends in oil production, discovery and reserves. The paper then summarises the key findings of the assessment in relation to decline rates and depletion rates (see Paper 9), methods of estimating resource size (see Paper 8), methods of forecasting future oil supply [76], the global resource base [29] and the future outlook for global oil supply (see Paper 6).

The approach is balanced and cautious, owing to the complexity of the relevant phenomenon, the multiple contributing factors, the serious limitations of the available data and the consequent uncertainties about future supply. The paper criticises several of the arguments of ‘peak oil’ advocates such as Campbell and Laherrère. For example, it concludes that global proved reserves are more likely to be underestimated rather than overestimated; so-called reserve growth is real, significant and not primarily the result of conservative reporting (see
Paper 9); the influential study of global resources by the US Geological Survey (USGS) [77] is well-founded and has not been discredited; trends in discovery and reserve growth since 2000 are consistent with the USGS estimates; and that widely used ‘curve-fitting’ techniques are unreliable and commonly lead to underestimates of recoverable resources and excessively pessimistic projections of future supply (see Papers 8 and 9).

At the same time the paper contests the conventional wisdom regarding future global oil supply which anticipates steadily increasing production up to 2030 [78]. The paper shows how, under a wide range of assumptions about the size of the global resource and the shape of the future production profile, a peak in the global production of conventional oil can be expected before 2030. It argues that forecasts that delay this peak until after 2030 rest upon a series of assumptions that appear to be at best optimistic and at worst implausible. This leads to the headline conclusion that a peak of conventional oil production before 2030 appears likely and there is a significant risk of a peak before 2020.

Since the assessment was published an increasing number of commentators have warned about medium-term constraints on global oil supply [e.g. 79, 80]. The IEA has continued to downgrade its long-term oil supply projections and now acknowledges that global production of crude oil is past its peak.

**Paper 7 - Oil futures: A comparison of global supply forecasts**

This paper compares and evaluates fourteen contemporary forecasts of the global supply of conventional oil and provides some observations on their relative plausibility. The first part of the paper clarifies some relevant concepts and definitions, summarises global trends in production, discoveries and reserves, and uses straightforward curve-fitting to demonstrate the insensitivity of the estimated date of peak production to the assumed size of the global resource. The paper then provides a detailed comparison of fourteen contemporary forecasts of global conventional oil supply over the period to 2030. Each of these forecasts is produced by a different energy-economic model and nine of them predict a peak in conventional oil production before that date.

The comparison is hampered by the lack of transparency in the relevant modelling tools, inconsistencies in the definition and coverage of different liquids, the wide range of approaches used and the multiplicity of assumptions made. Despite this, the paper shows how the forecasts can be usefully compared along two dimensions, namely: the shape of the future production profile (in particular the rate of production decline following the peak) and the assumed or implied ultimately recoverable resource of conventional oil. Other differences

---

9 Excluding natural gas liquids, and non-conventional sources such as oil sands and biofuels.
between the forecasts are either secondary or are components of these two parameters. ‘Quasi-linear’ forecasts such as those from the IEA are shown to imply global recoverable resources that are comparable to or greater than the mean estimate of the USGS, as well as a relatively rapid decline in conventional production following the peak (>3%/year).

The paper then assesses the implications of the different forecasts for the future rates of discovery, reserve growth and depletion\(^\text{10}\) and compares this to historical experience at the regional and global level (derived from our own analysis of industry data). As indicated above, this analysis suggests that forecasts which delay this peak until after 2030 rely upon a series of assumptions that appear at best optimistic and at worst implausible. For example, the reference scenario of the IEA WEO 2010 implies an average rate of depletion of global conventional oil resources that is several times greater than the maximum rate previously achieved in any oil producing region anywhere in the world. The paper therefore concludes that a peak of conventional oil production before 2030 is likely. It argues that short to medium term production prospects are more difficult to assess since they are more sensitive to political factors, the level of upstream investment and the complex and poorly modelled interactions between supply and demand – the most likely outcome of which is to turn a sharp peak into a ‘bumpy plateau’. Nevertheless, the analysis suggests there is a significant risk that conventional oil production will enter sustained decline before 2020.

I was invited to propose this paper as a candidate for the Eni Award 2012.

**Paper 8 - Hubbert's legacy: A review of curve-fitting methods to estimate ultimately recoverable resources**

This paper explores how logistic and other curves may be statistically fit to historical trends in oil production and discovery in a region and projected forward in time to estimate the ultimately recoverable resources (URR) of the region. These techniques were pioneered by Hubbert [81, 82] and are used by many analysts concerned about ‘peak oil’. This paper classifies and explains these techniques and identifies both their relative suitability in different circumstances and the level of confidence that may be placed in their results.

The paper first discusses the interpretation and importance of URR estimates, indicates the close relationship between curve-fitting and other methods of estimating URR and identifies three groups and nine individual types of curve-fitting technique. It then investigates each group in turn, indicating their historical origins, contemporary application and major strengths and weaknesses.

\(^{10}\) Defined as the rate at which the recoverable resources of a field or region are being produced.
The paper shows how curve-fitting implies a skewed field size distribution and diminishing returns to exploration, with the large fields being found early. But these assumptions only hold if depletion outweighs the effect of technical change, the region is geologically homogeneous and exploration has been relatively uninterrupted - which is rarely the case. The paper also shows how most applications of curve-fitting take insufficient account of its numerous weaknesses, including the sensitivity of the estimates to the choice of functional form, the risk of ‘over-fitting’, the inability to anticipate future cycles of production or discovery, the uncertainty in reserve estimates, the need to adjust these estimates to allow for future reserve growth and the effect of economic and political variables. In combination, these weaknesses are likely to contribute to underestimates of the URR.

The paper uses data on production and discovery from seven oil-producing regions to investigate the consistency of URR estimates from curve-fitting techniques and hence the level of confidence that can be placed in their results. By repeated application of non-linear regression, it demonstrates the unreliability of these techniques when (as is usually the case) they are applied at the country level with data that has not been corrected for future reserve growth. It shows how different techniques, together with variations in the length of time series, functional form and number of curves, repeatedly lead to inconsistent results - even for well-explored regions that are well past their peak of discovery and production [for details, see 29]. The paper also highlights some of the statistical issues raised and suggests how they may be addressed.\(^{11}\)

The paper concludes that the applicability of curve-fitting techniques is more limited than adherents claim and that the confidence bounds on the results are wider than commonly assumed. However, the degree of uncertainty declines as exploration matures, so the accuracy of regional and global URR estimates should improve over time. In addition, some of the limitations of curve-fitting may be overcome with the use of hybrid models that incorporate relevant economic and political variables. But data on these variables is rarely available, and despite providing a better fit to historical trends such models may not lead to substantially different estimates of the URR.

Since this paper was published, an increasing number of studies have applied curve-fitting techniques to different hydrocarbon resources [83-86]. Although some of these studies recognise the limitations identified above, most continue to make the same errors.

\(^{11}\) A paper providing a more comprehensive summary of the curve-fitting analysis will be published in a forthcoming edition of the *Philosophical Transactions of the Royal Society A*. 
Paper 9 - Shaping the global oil peak: A review of the evidence on field sizes, reserve growth, decline rates and depletion rates

This paper summarises and evaluates the evidence regarding four variables that have a critical influence on future global oil supply. The paper shows that, despite serious data limitations, knowledge of these issues has improved over the past decade and points to significant risks for future global supply.

The paper first evaluates the evidence regarding the size distribution of oil fields and highlights the critical importance of small number of ‘giant’ fields. Most of these fields are relatively old, many are well past their peak of production and most of the rest will begin to decline within the next decade or so. While technical improvements and higher prices should make more small fields viable, many will remain uneconomic and the exploitation of the rest will be subject to rapidly diminishing returns.

The paper then shows that reserve growth is real, significant and not primarily the result of conservative reserve reporting. Using industry data, it estimates the global average reserve growth since 1995 and shows that this is broadly in line with the optimistic assumptions made by the USGS - although most derives from countries with the largest reserves and poorest quality data. The paper also finds that reserve growth is greater for larger, older and onshore fields, so as global production shifts towards newer, smaller and offshore fields the rate of reserve growth should decrease in both percentage and absolute terms.

Comparing three studies that utilise confidential data on individual fields, the paper estimates that the production-weighted global average decline rate of post-peak fields is at least 6.5%/year and the corresponding decline rate of all currently producing fields is at least 4%/year. Both are on an upward trend as more giant fields enter decline, as production shifts towards smaller, younger and offshore fields and as changing production methods lead to more rapid post-peak decline. This implies that more than two thirds of current crude oil production capacity may need to be replaced by 2030, simply to maintain current levels of production.

Using a mix of industry and public domain data, the paper then estimates historical ‘depletion rates’ for oil-producing regions – namely, the ratio of annual production to the estimated ultimately recoverable resources. It shows that the maximum observed depletion rate at the regional level is ~5%/year and that the average depletion rate is much less. It also finds that most giant fields and most countries have reached their peak well before half of their recoverable resources have been produced. The paper then argues that ‘quasi-linear’ global supply forecasts such as those from the IEA are inconsistent with this historical experience and
also rely upon optimistic assumptions about the future rate of production decline from post-peak fields. But slower depletion or more rapid decline will lead to an earlier global peak.

Further Research

The synthesis reports for both assessments identified a wide range of promising topics for further research [12, 23]. In the case of rebound effects, the priorities include rigorous experimental studies of the impact of household energy efficiency improvements, the extension of econometric estimates to a wider range of sectors and energy services, the use of CGE models to systematically investigate economy-wide rebound effects, and econometric analysis of the links between energy and total factor productivity. In the case of global oil depletion, the priorities include: using industry databases to explore the nature, scale and determinants of reserve growth; evaluating the technical and economic potential for enhanced oil recovery; developing modelling tools that integrate the physical and economic determinants of oil supply and demand; exploring the availability, cost and lead times for developing non-conventional resources; quantifying the current and expected future ‘net energy yield’ of liquid fuel supply; and appraising the economic and technical potential of various demand side mitigation options.

I am seeking to pursue several of these research topics and have successfully secured funding for two follow-up projects. The first (funded by ESRC/Defra) is quantifying direct and indirect rebound effects for households through a combination of environmentally-extended input-output modelling and the econometric analysis of household survey data. The former is used to estimate the embodied energy of different categories of household goods and services, while the latter is used to estimate expenditure, own-price and cross-price elasticities [87]. One journal paper has been published [88] and two are in submission.

The second (funded by the Joint Research Centre of the European Commission) is using systematic review techniques to assess and compare the available estimates on regional and global shale gas resources and to evaluate the underlying resource assessment methodologies. Two journal articles are in preparation.

In addition, I have two research proposals in submission (on rebound effects and net energy yields respectively), and the special edition of Philosophical Transactions A that I am editing has invited contributions on several of the above topics.
References


http://sepwww.stanford.edu/sep/jon/world-oil.dir/lynch/worldoil.html


Turner, K. *A computable General equilibrium analysis of the relative price sensitivity required to induce rebound effects in response to an improvement in the energy efficiency of the UK economy*; Department of Economics, University of Strathclyde: Glasgow, 2008.


82. Hubbert, M. K., Techniques of prediction as applied to the production of oil and gas; Oil and gas supply modeling; proceedings of a symposium. *Oil and gas supply modeling, Washington, DC, United States, June 18-20, 1980* 1982, 16-141.


Part B

Analysing controversies in energy policy: Assessing the evidence for rebound effects and global oil depletion

The following papers are included within Part B:

|---|---|
Improving the evidence base for energy policy: The role of systematic reviews

Steve Sorrell*

Sussex Energy Group, SPRU (Science and Technology Policy Research), Freeman Centre, University of Sussex, Falmer, Brighton BN1 9QE, UK
Available online 1 August 2006

Abstract

The concept of evidence-based policy and practice (EBPP) has gained increasing prominence in the UK over the last 10 years and now plays a dominant role in a number of policy areas, including healthcare, education, social work, criminal justice and urban regeneration. But despite this substantial, influential and growing activity, the concept remains largely unknown to policymakers and researchers within the energy field. This paper defines EBPP, identifies its key features and examines the potential role of systematic reviews of evidence in a particular area of policy. It summarises the methods through which systematic reviews are achieved; discusses their advantages and limitations; identifies the particular challenges they face in the energy policy area; and assesses whether and to what extent they can usefully be applied to contemporary energy policy questions. The concept is illustrated with reference to a proposed review of evidence for a ‘rebound effect’ from improved energy efficiency. The paper concludes that systematic reviews may only be appropriate for a subset of energy policy questions and that research-funding priorities may need to change if their use is to become more widespread.

Keywords: Evidence-based policy and practice; Systematic reviews; Research synthesis

1. Introduction

From relatively small beginnings within the medical field, the concept of evidence-based policy and practice (EBPP) has gained increasing prominence in the UK over the last 15 years (Solesbury, 2001). An infrastructure now exists for developing and promoting evidence-based medicine and healthcare, both in the UK and abroad,¹ and the concept has been adopted by other professional and policy areas, including education, social work, criminal justice and urban regeneration (Davies et al., 2000).² The UK Economic and Social Research Council has established a network for conducting evidence-based reviews in the social sciences and is funding a Centre for EBPP at Queen Mary College, University of London. The concept informed the 1999 White Paper on Modernising Government, together with a number of subsequent UK government publications (Bullock et al., 2001), and is central to the remit of the Cabinet Office. But despite this substantial, influential and growing activity, EBPP appears to be largely unknown to policymakers and researchers within the energy field (Gross, 2005).

This paper aims to introduce the concept of EBPP to energy policy researchers and evaluate its potential contribution to research and practice in this area.³ Specifically, the paper seeks to:

- propose a definition of EBPP and identify its key features;

---

¹Including the Cochrane Collaboration, the National Institute for Clinical Excellence, the Health Development Agency and the NHS Centre for Reviews and Dissemination.

²These have established parallel initiatives such as the Centre for Evidence Informed Education Policy and Practice (EPPI) at the University of London (Davies et al., 2000).

---

0301-4215/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.enpol.2006.06.008
• introduce the concept of a systematic review of evidence in a particular area of policy or practice, and summarise the methods through which this is achieved;
• discuss the advantages and limitations of systematic reviews and the particular challenges they face in the energy policy area; and
• assess whether and to what extent the systematic review methodology can usefully be applied to contemporary energy policy questions.

The following section introduces EBPP, identifies its different elements and highlights the disputed nature of ‘evidence’ in different policy areas. Section 3 describes a systematic review of evidence and summarises the stages and methods through which this is achieved. Section 4 highlights the limitations of systematic reviews, focusing upon a number of interrelated ‘biases’ to which they are prone. Section 5 discusses the challenges of applying systematic reviews to energy policy questions and illustrates this with reference to an example. Section 6 concludes.

2. What is EBPP?

The rhetoric of EBPP pervades many areas of policy and practice within the UK. As the title suggests, its influence extends from low-level issues of professional practice, such as classroom techniques to encourage pupils with behavioural difficulties (Evans and Benefield, 2001), to high-level policy issues, such as the relationship between poverty, income inequality and violence (Hsieh and Pugh, 1993). But as with most such concepts, the precise meaning of ‘evidence-based’ is contested.

EBPP has been defined as ‘the integration of experience, judgement and expertise with the best available external evidence from systematic research’ (Davies, 1999). Hence, EBPP implies striking a balance between formal research evidence and professional judgement. Davies and Nutley (2002) argue that for evidence to have a greater impact, there must be agreement as to what counts as evidence in what circumstances, a strategic approach to the creation of evidence in priority areas, systematic efforts to accumulate and synthesise evidence, effective dissemination of evidence to where it is most needed and initiatives to ensure better utilisation of evidence. But while each of these may be a necessary requirement for EBPP, the first presents a particular challenge. Most applications of EBPP have prioritised scientific research over other forms of evidence and given greater weight to some research methodologies than to others. Box 1 identifies four general types of research methodology, listing them in descending order of their weighting in EBPP.

When EBPP is extended to other areas of policy and practice where (quasi-) experimental studies are less feasible or appropriate, the consensus over the relative weight to give to different research methodologies can break down. Table 1 provides a stylised distinction between those areas of policy and practice where a consensus broadly exists and those where it does not. In practice, the situation may be better represented as a spectrum of possibilities, with disputes arising even within “core” areas of EBPP such as healthcare. Furthermore it is possible to adopt widely different methodological approaches to very similar research questions.

These distinctions are important when assessing the potential role of EBPP within energy policy research. For example, where on this spectrum does energy policy research, or different approaches to this research, lie? Table 2 compares the methodological approaches in energy policy research with those taken in other policy areas (Davies et al., 2000).

EBPP practitioners have identified a number of problems with the existing systems for creating, synthesising, disseminating and utilising research evidence and have provided a number of suggestions on how these may be improved. But the greatest weakness lies in the inadequate accumulation and synthesis of research results and the poor quality of traditional literature reviews. This is a dominant theme within EBPP and the primary rationale for developing systematic reviews of the available evidence base.

3. The methodology of systematic reviews

There are several reasons why systematic reviews have become so central a feature of EBPP. First, experience in the medical field and elsewhere suggests that policy and practice are often based on inadequate evidence. For example, the UK Department of Health (1991) found that only around 15% of the clinical interventions used in the National Health Service were based upon unequivocal scientific evidence and many commonly used interventions did not improve the health of patients (some, indeed, appeared to make their condition worse). In addition, interventions that were well supported by evidence were often neglected or ignored. Second, the increasing volume of research, or different approaches to this research, lie? Table 2 compares the methodological approaches in energy policy research with those taken in other policy areas (Davies et al., 2000).

EBPP practitioners have identified a number of problems with the existing systems for creating, synthesising, disseminating and utilising research evidence and have provided a number of suggestions on how these may be improved. But the greatest weakness lies in the inadequate accumulation and synthesis of research results and the poor quality of traditional literature reviews. This is a dominant theme within EBPP and the primary rationale for developing systematic reviews of the available evidence base.

3. The methodology of systematic reviews

There are several reasons why systematic reviews have become so central a feature of EBPP. First, experience in the medical field and elsewhere suggests that policy and practice are often based on inadequate evidence. For example, the UK Department of Health (1991) found that only around 15% of the clinical interventions used in the National Health Service were based upon unequivocal scientific evidence and many commonly used interventions did not improve the health of patients (some, indeed, appeared to make their condition worse). In addition, interventions that were well supported by evidence were often neglected or ignored. Second, the increasing volume

4. In some areas (most notably healthcare) the need for evidence and the nature of convincing evidence is a given. In other areas (most strikingly, social care), the very nature of evidence is hotly disputed and there is strong resistance to assigning privileged status to one research method over another. Such divergent attitudes arise from deep-rooted ontological and epistemological assumptions. Furthermore, where post-modern perspectives are prevalent there is a general distrust of any notion of objective evidence. (Davies et al., 1999)

5. Compare for example sociological approaches to household energy consumption (Hand et al., 2003) to econometric estimates of household production functions (Willet and Naghshpour, 1987).

6. The same problem is found in other areas. For example, prison tour programmes for preventing juvenile delinquency have enjoyed a great deal of popularity, despite only anecdotal evidence of their success. A
Box 1
Different types of research evidence

- **Experimental and quasi-experimental evidence** is exemplified by randomised controlled trials in medicine, but also includes controlled before-and-after studies and various types of matched comparison that can be applied to policy evaluation and broader social scientific questions. This type of study potentially provides reliable evidence of the causal effect of different mechanisms by explicitly controlling for the effect of different variables.

- **Survey and econometric evidence** provides an alternative technique for exploring causal hypotheses, either through surveys conducted by the researcher or through the econometric analysis of existing data. In addition, survey data can provide valuable information about the nature, size, frequency and distribution of a particular variable or problem.

- **Modelling evidence** covers a variety of approaches to analysing the operation and consequences of different mechanisms using a simplified mathematical model. Like all theories, these models abstract from real-world complexities and focus on key mechanisms, either conceptually or by combining theoretical assumptions with empirical data.

- **Qualitative evidence** includes a variety of techniques for obtaining information regarding the opinions, attitudes and perceptions of individuals and groups in different contexts. Examples include case studies, participant observation and focus groups. Frequently, qualitative evidence will be combined with quantitative (e.g. survey) evidence to improve understanding about how and why a policy or mechanism works and under what conditions. Hence, rather than distinguishing between qualitative and quantitative approaches, it is perhaps more useful to compare those approaches that seek to establish causal inferences, whether through quantitative methods, qualitative methods or both, as compared to those which seek to interpret the experience of individuals and identify the meanings which those experiences hold. The former tends to be associated with economics, political science and social psychology, while the latter tends to be associated with sociology and social anthropology.

Table 1

<table>
<thead>
<tr>
<th>Issue</th>
<th>‘Consensus’ area of policy research</th>
<th>‘Disputed’ area of policy research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcomes of policies and interventions</td>
<td>Simple</td>
<td>Complex</td>
</tr>
<tr>
<td>Objectives of policies and interventions</td>
<td>Simple, reinforcing and consensual</td>
<td>Multiple, competing and contested</td>
</tr>
<tr>
<td>Dominant research methodology</td>
<td>Experimental and quasi experimental</td>
<td>Qualitative methods</td>
</tr>
<tr>
<td>Degree of consensus regarding appropriate research methodology</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Importance of context in determining outcomes</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Emphasis given to evaluating policies and interventions</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Typical objective of such evaluations</td>
<td>Whether the policy/intervention works</td>
<td>Why, how and in what circumstances the policy/intervention works</td>
</tr>
</tbody>
</table>

of research findings makes it difficult for policy makers and practitioners to keep abreast of current understanding, creating a need for more effective synthesis of research results. Third, a combination of the complexity of the relevant issues, the variable quality of research evidence and the methodological and other biases of individual researchers, leads to conflicting recommendations by different authors and corresponding uncertainty over whom to trust. This problem can be exacerbated by the selective use of evidence by powerful interest groups and by the partial and unbalanced treatment of research results by the media. Finally, the traditional form of research

(footnote continued)

systematic review of the subject found that the programmes had the opposite effect from that intended—they made crime more likely (Petrosino et al., 2000).

*King et al. (1994) provide an excellent account of the use of qualitative methods for investigating causal hypotheses.
synthesis—termed the narrative review by EBPP practitioners—often fails to address these limitations, with the result that different reviewers can reach different conclusions from the same research base.\(^7\) More specifically, the criticisms levelled at traditional narrative reviews include (Petticrew and Roberts, 2005):

- Poor specification of the review topic, leading to excessively wide-ranging discussion and inconclusive results.
- Selective and opportunistic use of evidence, leading to selection bias and the neglect of relevant studies.
- Inadequate specification of the criteria for including or excluding studies from a review.
- Limited attention to methodological quality, leading to a lack of discrimination between sound and unsound studies.
- Lack of transparency, encouraging subjectivity and bias in the reporting of results.

Both research practitioners and commentators on EBPP have disputed these criticisms (Hammersley, 2001), but there is evidence to suggest that they are valid. For example, Oliver et al. (1999) compared six different narrative reviews of accident prevention for older people, all conducted between 1995 and 1999 and covering a total of 137 studies. Of these, only 33 studies were common to at least two reviews, only two were common to all six reviews and only one study was treated consistently by all six. Similarly, Oakley and Fullerton (1995) found 70 studies of smoking prevention programmes for young people, of which only 27 had been captured by two previous narrative reviews. Of these, only three were common to both reviews and the differences in coverage led the two reviews to reach different conclusions on the relative effectiveness of different prevention programmes.

Systematic reviews of existing evidence seek to address each of the above criticisms. Ideally, the systematic review will reach an authoritative conclusion regarding a specific research question through a comprehensive synthesis of currently available primary research. But where this is not possible, the review will seek to explain the reasons for the differences between studies and to identify priority areas for further research.

Systematic reviews use explicit and transparent methodologies that are replicable and updateable. They involve:

- A clear specification of the research question(s) to be addressed.
- Systematic and exhaustive searching of the available literature.
- Applying explicit criteria for the inclusion or exclusion of studies.
- Appraising the quality of the included studies using transparent and standardised criteria.
- Summarising and synthesising the results in an ‘objective’ manner.
- Disseminating the results effectively to the appropriate audience.
- Updating the results of the review at intervals, when new research becomes available.

---

\(^7\)Oakley (2002) is particularly disparaging: ‘...Most traditional literature reviews are discursive rampages through selected bits of literature that the researcher happens to know about or can easily reach on his or her bookshelves at the time’.

---

Table 2
Methodological preferences and debates in different policy areas

<table>
<thead>
<tr>
<th>Policy area</th>
<th>Methodological preferences and debates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthcare</td>
<td>Gold standard of randomised controlled trials with additional methodological safeguards. Growing interest in qualitative methods to give a complementary view</td>
</tr>
<tr>
<td>Education</td>
<td>Competing rather than complementary methods with methodological disputes between different groups. Econometric analysis of large data sets, but little experimental research</td>
</tr>
<tr>
<td>Criminal justice</td>
<td>General acceptance of experimental or quasi-experimental methods</td>
</tr>
<tr>
<td>Social care</td>
<td>Preference for qualitative methodologies. Quantification and experimentation often viewed with suspicion and hostility</td>
</tr>
<tr>
<td>Welfare policy</td>
<td>Eclectic use of methods to provide complementary insights. Some longitudinal study but almost no experimentation</td>
</tr>
<tr>
<td>Housing</td>
<td>Predominant use of qualitative and quantitative survey methods. Use of econometric methods for forecasting housing needs. Recent emergence of more multi-disciplinary approaches</td>
</tr>
<tr>
<td>Urban policy</td>
<td>Difficulties in relating particular outcomes to particular interventions and in identifying externalities. Diverse methods employed, but particular reliance on case studies. Little or no experimentation</td>
</tr>
<tr>
<td>Transport</td>
<td>Multidisciplinary. Policy research frequently rooted in economic modelling and forecasting</td>
</tr>
<tr>
<td>Energy policy</td>
<td>Multidisciplinary, but with a strong bias towards econometric analysis of secondary data and economic modelling. Policy evaluation relatively weak in many areas</td>
</tr>
</tbody>
</table>

Source: Adapted from Davies et al. (2000). Text on energy policy added.
Prior to conducting a systematic review, the review team will publish their intended approach to each of these issues in a Protocol. This will describe the scope of the review, the strategy for searching the evidence base, the criteria for the inclusion and exclusion of evidence, the basis for assessing the quality of such evidence and the procedure for synthesising the results. The Protocol will be open to comment by relevant stakeholders, with the aim of improving transparency and credibility. Organisations such as the Campbell Collaboration have published guidelines for producing such Protocols, together with libraries of examples. The rigorous and comprehensive nature of a systematic review leads to correspondingly greater resource requirements than for a comparable narrative review, which may create difficulties for research funders.8

The differences between systematic reviews and traditional narrative reviews are summarised in Table 3. The following sections describe each stage of a systematic review in more detail.

3.1. Choosing the review questions

In contrast to traditional literature reviews, systematic reviews usually employ a sharply defined research question, or a hypothesis that is suitable for test through empirical research. The UK Cabinet Office (2005) recommends that these questions specify the mechanisms, interventions or policies in question; the population and/or subgroups in question; the outcomes that are of interest; and the context in which the question is set. For example, rather than reviewing the literature on ‘barriers to energy efficiency’, a systematic review should seek to answer a more specific question such as

What is the evidence that the inability to appropriate the cost savings from energy efficiency investment (mechanism) leads households in rental accommodation (population) to use energy less efficiently than owner occupiers (outcome)?

Review questions are normally refined to make them more specific and elaborated by specifying the objectives more clearly.9 In practice, systematic reviews tend to focus primarily on the effects of individual policies or interventions, as opposed to other types of mechanism, and to prioritise the use of (quasi-) experimental evidence, as opposed to other types of evidence. An example of a quasi-experimental study would be a comparison of energy consumption by participants in a demand-side management scheme with that of a control group of non-participants (Meyer, 1995).

8The EPPI Centre, for example, recommends as a minimum two researchers at 50% full-time equivalent for 9 months, together with additional resources for information scientists. Given the limited funding available for energy policy research (compared, for example, with healthcare), a greater use of systematic review methodologies would raise important questions of value for money and may prove impractical in the near term.

9For example, in a systematic review of the ‘effects of neighbourhood watch schemes on preventing crime’, the following objectives were established: (a) to operationalise the inputs (e.g. the policies qualifying as neighbourhood watch) and the outcomes (e.g. the relevant types and measures of crime); (b) to identify studies that evaluate the effect of the qualifying schemes on the relevant outcomes; (c) to identify a list of studies that meet minimum criteria of scientific rigour; (d) to obtain a comparable measure of effect size in the selected most rigorous studies; and (e) To arrive at a conclusion about the effectiveness of neighbourhood watch schemes (Bennett et al., 2004).
3.2. Searching the literature

Systematic reviews aim to be comprehensive in their coverage of the available literature, including both peer reviewed academic papers and ‘grey’ literature such as Ph.D. dissertations, conference papers and consultants’ reports. However, while non-peer reviewed literature is generally admissible, the absence of peer review may be one factor taken into account when assessing the quality of the studies. Similarly, while systematic reviews should not be restricted to English language sources, resource constraints often necessitate this in practice.

The Protocol specifies the databases, bibliographies, contacts and other sources that are to be employed; the years to be covered; the search strategies to be used (e.g. keywords); and the mechanisms available for retrieving less accessible documents. In practice, the relative importance of different information sources and search strategies will vary between different fields. Typically, the literature search is resource intensive and requires the support of information specialists. Also, many more studies are normally identified than are finally used in the review. For example, in reviewing the evidence on the effectiveness of policy measures to prevent homeowners defaulting on their mortgage, Wallace et al. (2004) initially identified a total of 1832 references. Of these only 766 were classified as generally relevant, 49 met the inclusion criteria and only 22 eventually passed the quality threshold.

3.3. Including and excluding studies

In parallel with defining specific questions, systematic reviews establish criteria for including and excluding studies from the review. For example, in the mortgage default study cited above, the inclusion criteria were that the study should: (a) address at least one of the four types of policy intervention defined in the protocol; (b) include assessment of policy effectiveness; and (c) be based upon empirical research (Wallace et al., 2004).

Applying criteria such as these can lead to the exclusion of a large number of the identified studies. For example, the mortgage study excluded 97% of the identified references, largely because they did not include empirical research. The exclusion of evidence on these grounds has been criticised by commentators who consider that expert opinion, theoretical models and other types of literature may be very relevant to particular research questions. In practice, many systematic reviews will use such studies as ‘conceptual background’ and will summarise the conclusions of these studies separately from the main text.

As with defining the research question, the choice of inclusion criteria involves a difficult compromise between comprehensiveness and quality control. If strict criteria are used, there is a risk that the majority of relevant studies will be excluded, but if loose criteria are used, synthesis will become more difficult.

3.4. Extracting information

Systematic reviews extract and store information from the included studies in a standardised way. Typical headings include the nature of mechanisms studied; the research methods used; the choice of population and sample; the methods for controlling intervening variables; the outcomes measured or observed; and the size of the effects. Several organisations provide standardised tools and databases for this purpose. Most of the standard frameworks are designed for quantitative studies, often with the aim of conducting a subsequent meta-analysis of results (see below). Similarly, most of the standard frameworks are designed for studies that evaluate the outcomes of particular policies or interventions and are poorly suited to other types of evidence.

The extraction process is normally conducted prior to the quality assessment of individual studies, which means that much of the extracted information may be discarded when drawing the final conclusions. However, the review database can provide a valuable reference source for the status of research in a particular area, and can make it easier for researchers to compare the results of different studies and to identify gaps.

3.5. Appraising quality

Explicit appraisal of the quality of individual studies is perhaps the defining feature of a systematic review, but also the most controversial. Most reviews use a weighted or unweighted checklist of requirements against which each study is scored. Only studies that exceed a certain quality threshold are retained for use in the final synthesis. A range of quality criteria has been developed, with most having a bias towards (quasi-) experimental studies. Three standard concepts that dominate are:

- **Internal validity**: The extent to which the study shows a cause-effect relationship between the independent and dependant variables.
- **Construct validity**: The adequacy of the operational definition and measurement of the theoretical constructs.
- **External validity**: The ‘generalisability’ of the proposed causal relationships among different actors, places and times.

---

10For example, indexed and comprehensive electronic databases are more established in healthcare than in most areas of social science. Keyword searching is also more difficult in the social sciences, owing to the relative imprecision and inconsistency in the use of language.

11Including the Campbell Collaboration and the Centre for Evidence Informed Education Policy and Practice at the University of London.
These may be refined into very specific criteria if a single type of quantitative evidence dominates, but may need to be stated in more general terms if multiple sources of evidence (including qualitative studies) are to be included. Table 4 lists the criteria used to appraise the quality of both quantitative and qualitative studies in the systematic review of ‘mortgage safety nets’ cited earlier (Wallace et al., 2004). This list demonstrates that quality appraisal is necessarily a matter of judgement (e.g. just when can a description of context be considered ‘adequate’?). However, the importance of subjectivity may vary with the type of study and the transparency of the appraisal process can be valuable in itself. Most systematic reviews seek to minimise subjectivity by having two or more researchers conduct quality appraisals and compare results.

The need for dedicated criteria to appraise qualitative research has become more pressing as EBPP techniques are extended to a wider range of policy areas. In response, numerous review teams have developed checklists of criteria, often with considerable differences in approach. In an attempt to introduce some standardisation, the Cabinet Office Strategy Unit has published a comprehensive review of criteria for assessing qualitative research and has proposed a synthesising framework involving 18 appraisal questions and 64 indicators (Spencer et al., 2003). More recent approaches have sought to extend appraisal beyond issues of methodological quality to include the quality of reporting and relevance to important policy questions (Boaz and Ashby, 2003). This recognises that excessive concentration on methodological quality may create the risk of including rigorous, but narrowly focused studies that are of little value to stakeholders, while at the same time excluding broader studies that are highly relevant to policy but have a weaker methodology. But since this proposal departs from one of the core tenets of systematic reviews—giving priority to methodologically rigorous studies—it has encountered resistance.

Table 4
Criteria for appraising the quality of both quantitative and qualitative studies

<table>
<thead>
<tr>
<th>Area</th>
<th>Issue</th>
<th>Essential/Desirable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>Is the research question clear?</td>
<td>Essential</td>
</tr>
<tr>
<td>Theoretical perspective</td>
<td>Is the theoretical perspective of the author(s) explicit and has this influenced the study design, methods or findings?</td>
<td>Desirable</td>
</tr>
<tr>
<td>Study design</td>
<td>Is the study design appropriate to answer the question?</td>
<td>Essential</td>
</tr>
<tr>
<td>Context</td>
<td>Is the context or setting adequately described?</td>
<td>Desirable</td>
</tr>
<tr>
<td>Sampling</td>
<td>Qualitative: is the sample adequate to explore the range of subjects and settings and has it been drawn from an appropriate population?</td>
<td>Essential</td>
</tr>
<tr>
<td></td>
<td>Quantitative: is the sample size adequate for the analysis used and has it been drawn from an appropriate population?</td>
<td>Essential</td>
</tr>
<tr>
<td>Data collection</td>
<td>Was the data collection adequately described and rigorously conducted to ensure confidence in the findings?</td>
<td>Essential</td>
</tr>
<tr>
<td>Data analysis</td>
<td>Was there evidence that the data analysis was rigorously conducted to ensure confidence in the findings?</td>
<td>Essential</td>
</tr>
<tr>
<td>Reflexivity</td>
<td>Are the findings substantiated by the data and has consideration been given to any limitations of the methods or data that may have affected the results?</td>
<td>Desirable</td>
</tr>
<tr>
<td>Generalisability</td>
<td>Do any claims to generalisability follow logically, theoretically and statistically from the data?</td>
<td>Desirable</td>
</tr>
<tr>
<td>Ethics</td>
<td>Have ethical issues be addressed and confidentiality respected?</td>
<td>Desirable</td>
</tr>
</tbody>
</table>

Source: Croucher et al. (2003).

---

12 As an illustration, a review of quasi-experimental studies in criminology used the following classification scheme (in ascending order of quality): (1) correlation between a prevention programme and a measure of crime at one point in time; (2) measures of crime before and after the programme, with no comparable control condition; (3) measures of crime before and after the programme in experimental and comparable control conditions; (4) measures of crime before and after the programme in multiple experimental and control units, controlling for other variables that influence crime; and (5) random assignment of programme and control conditions to units.

13 This need has also become apparent within ‘core’ EBPP applications such as healthcare. For example, a systematic review of the influences on the take-up of childhood immunisations found that several key determinants were missed if only quantitative studies were included (Roberts et al., 2002).

14 For example, Oakley (2002) compared the criteria proposed by four different review teams: out of a total of 46 criteria, 28 occurred in only one of the lists, 10 in two and six in three, with only two criteria being common to all four lists.
3.6. Synthesising results

The final stage of a review (excluding dissemination and updating) is the synthesis of results from the subset of studies that pass the quality appraisal. If these are quantitative studies, the synthesis can be conducted in a rigorous way, using a group of statistical techniques known as meta-analysis (Cooper and Hedges, 1994). The objective may either be to obtain a more accurate estimate of the existence and size of the relevant effect, or to explain the variations in results between studies. By combining the results of several studies, the overall sample size may be increased and statistical tests may have a greater probability of detecting the relevant effects (i.e. greater power).

Meta-analysis would appear to be particularly suitable for the synthesis of certain types of econometric data, such as estimates of price or income elasticities. To date, however, the level of take-up within the economics profession appears limited (Florax et al., 2002). Meta-analysis is much less suitable when there are significant differences in the mechanisms or policies under investigation, the outcomes measured, or the context in which the studies took place. In these cases, a narrative synthesis is likely to form an important part of any review, but the EBPP literature provides much less guidance on how this should be achieved. Where quantitative and qualitative research is combined there is a tendency to prioritise the first and use the second to aid the interpretation of the quantitative findings.

4. The limitations of systematic reviews

Systematic reviews have been criticised by researchers and practitioners in a number of fields (Hammersley, 2001). Some of these criticisms are misconceived and have been effectively countered by EBPP advocates (see Box 2), while others have their roots in long-running disputes within the philosophy of science. These criticisms have been applied to the original application of systematic reviews—experimental studies of highly specific medical interventions—but become more vigorous when the approach is extended to complex policy issues. Research in these areas is frequently qualitative and/or characterised by disputes between competing disciplines.

In this context, it is unsurprising that the assumptions of systematic reviews should encounter difficulties. What is less clear is whether, as a result, the approach has nothing to offer policymakers and researchers, or whether it can be usefully adapted to address interesting and relevant policy questions. This section highlights three interrelated ‘biases’ in the systematic review methodology, namely:

- bias in the selection of question;
- bias in the selection and appraisal of evidence; and
- bias in the synthesis of results.

4.1. Bias in the selection of question

Systematic reviews require clear, specific and answerable questions that are tightly and narrowly defined. At the same time, the questions should be relevant to current debates and of sufficient significance to justify the resources spent. But there is a tension between these objectives: narrow questions may be more answerable but of less interest to policymakers and practitioners (especially given the resource requirements of a systematic review), while broad questions may be more interesting but less amenable to the systematic approach.

Systematic reviews commonly address ‘micro’ questions regarding the technical efficiency and effectiveness of particular policies or practices, but they rarely address ‘macro’ policy questions, or those with a substantial normative content. Since such questions are relevant to all areas of policy, the bias results more from the constraints of the methodology, than from the characteristics of the policy sectors to which it has been applied.

Most systematic reviews formulate a question of the form: ‘What is the effect of policy/intervention X in population Y on outcome(s) Z?’ This focus on ‘what works’ reflects both the medical origins of EEBP and its strong links with policy evaluation. But ‘what works’ is neither the only question of interest to policymakers or

Box 2
‘Myths’ about systematic reviews (Source: Petticrew, 2001)

- **Systematic reviews are the same as ordinary reviews, only bigger**: While they are more comprehensive, features such as specificity, transparency and quality appraisal also make them qualitatively different.
- **Systematic reviews include only randomised controlled trials**: A variety of quantitative and qualitative methodologies have been included within systematic reviews.
- **Systematic reviews are of no relevance to the real world**: Systematic reviews have been used to examine a range of contemporary and contentious policy issues such as domestic violence and child abuse.
- **Systematic reviews necessarily involve statistical synthesis**: Many systematic reviews do not use statistical methods.
- **Systematic reviews are a substitute for individual studies**: Systematic reviews do not necessarily provide definitive answers and often identify the need for additional primary research.
necessarily the most important question. For example, other relevant questions may include: What are the recent trends in \( X \)? What are the causes of \( X \)? What are the risks of \( X \)? What are the likely costs and benefits of \( X \)? What might happen if \( X \) is done? (Solesbury, 2001, p. 8).

A useful distinction here is between *descriptive* questions—such as: ‘what are the recent trends in UK industrial fuel consumption?’ and *causal* questions—such as: ‘what are the reasons for those trends? Both provide useful information for policymaking and both involve theoretical assumptions—for example, in defining measurement categories or in choosing between decomposition methodologies. But historically, systematic reviews have focused almost exclusively on the latter.

A possible way of subdividing causal questions is to distinguish between *historical* questions, which enquire into the reasons for present or historical states of the world, and *forecasting* questions, which estimate future states of the world under given assumptions. Both are relevant to policymaking, but systematic reviews have focused almost exclusively on the former.

The sources of evidence for causal questions include both *theoretical* studies that postulate, model or explain relationships between variables and *empirical* studies that calibrate or test those frameworks using historical data. Both are relevant to causal questions, although individual studies may be either largely theoretical or largely empirical, as well as either largely quantitative or largely qualitative. But systematic reviews have focused primarily on empirical studies, and particularly those that use quantitative methodologies.

These distinctions (summarised in Table 5) are relevant to the potential application of systematic reviews to energy policy research. While systematic reviews have traditionally focused on historical questions, energy policy researchers are frequently preoccupied with forecasting questions such as the technical and economic potential of particular technologies. Similarly, while systematic reviews have traditionally focused on empirical questions, such as evaluating the success (under some measure) of particular policies, energy policy researchers have spent a great deal of time on theoretical questions. The distinction is relative rather than absolute, since all empirical questions assume a particular theoretical framework and individual studies may test, develop or modify these frameworks as appropriate. Nevertheless, systematic reviews have tended to assume that there is broad consensus on the appropriate framework and hence provide relatively little guidance on how to adjudicate between competing theoretical approaches.

Energy policy research does share common ground with systematic reviews in that priority tends to be given to quantitative methodologies. However, systematic reviews have a bias towards (quasi-) experimental studies while energy research has a bias towards econometric analysis of secondary data and various forms of economic modelling. These models are based in varying degrees upon real-world data, but also embody a large number of theoretical assumptions. While a systematic comparison of modelling results has been undertaken (Barker et al., 2002), there appears to be little precedent for this within the systematic review literature.

### 4.2. Bias in the selection and appraisal of evidence

Systematic reviews were originally applied to ‘what works’ questions within medicine. These are best addressed through experimental studies, utilising explicit and replicable procedures that allow for physical or statistical control of intervening variables, together with the generalisation of results. This notion of a ‘gold standard’ methodology has influenced the application of systematic reviews in all areas, even where experimental studies are less feasible. As a result, this approach has been criticised for its allegedly ‘positivist’ assumptions, including: first, that society should be studied in the same way as the natural world; second, that causality may be established by identifying empirical regularities among sequences of events; third, that some research methodologies are better than others in all instances; and fourth, that subjectivity is a source of bias that should be minimised (Hammersley, 2001).

Each of these assumptions has been the subject of sustained criticism within the philosophy of science for many years. Some practitioners reject all of them and argue that the appropriate aim of social science is to elucidate meanings. But most take a compromise position, emphasising the differences between natural and social science but still claiming that causality has meaning in social systems and can be identified in particular instances. However, since the social and natural worlds are fundamentally different, there are likely to be comparable differences in the methodologies that can be employed and the status of the knowledge claims that can be made (Sayer, 1992).

In natural science, causality is normally established through identifying regularities in the outcomes of repeated experiments. But doing the same for public policies or complex social interventions can be problematic. Regularities are achievable in natural science because experiments establish ‘closed’ systems where the mechanism possessing the causal power is stable, and where the external conditions in which the mechanism is situated are held constant (Bhaskar, 1975). These conditions are practically impossible to achieve in social situations because: first, the
relevant mechanisms involve the actions of people who can reason, learn and change their behaviour; and second, there are a host of complex contextual influences that are difficult to control. As a result, the relationship between particular causes and effects is likely to vary across time and space. The same causal mechanism may produce different outcomes according to context, the same outcome may be produced by different causal mechanisms, and the operation of different mechanisms may not be stable since they depend upon the understanding and behaviour of individuals. These differences have led some authors to argue that the scope for ‘generalisation’ within social science is limited and the best that can be hoped for is the identification of partial regularities that hold for only a limited period of time (Lawson, 1997). Hence, the fact that a particular relationship between cause and effect does not hold in a particular instance may not be a reason to reject the underlying theory, because a host of offsetting contextual influences may be at work.

This stance leads the critics of ‘positivist’ social science to prefer qualitative rather than quantitative research methods, and to emphasise the theoretical content of explanations as much as the statistical relationships between observed variables (Sayer, 1992). Rival theories are then judged not solely in terms of their empirical success but also in terms of their ‘explanatory adequacy’ (Lawson, 1997). However, this preference has been challenged by Pratschke (2003) and others who argue that statistical approaches such as econometrics are entirely compatible with this more nuanced understanding of the nature of causality in social systems (Bache, 2003; Hodgson, 2004). This is largely because there are reasons to expect relatively enduring regularities in some social systems, and methods such as econometrics can reveal such regularities while controlling as far as possible for intervening variables:

... although the social world is potentially, or logically, a purely open system, in practice it is characterised by ‘quasi-closures’ which may manifest themselves in stable patterns of events…Faced with the challenge to make decisions in complex situations, agents are inclined to develop habits, conventions and routines. Such habitual and routine behaviour will be reflected in social institutions which may prove to be stable over periods of time. (Bache, 2003, p. 14).

This methodological debate is ongoing. But one reasonable conclusion is that the research methods should be appropriate to the subject matter under investigation—with no one method being necessarily superior to any other. While some EBPP authors accept this, there is still a tendency to prioritise (quasi-) experimental studies within a ‘hierarchy’ of methodologies (Hammersley, 2001). Qualitative studies are often given a ‘supporting role’, either providing mere ‘conceptual background’ or being summarised separately from the main meta-analysis. But the contemporary philosophy of science provides little justification for this and it can present practical difficulties when systematic reviews are applied to questions where qualitative research (quite appropriately) dominates. As a result, Boaz and Ashby (2003) have advocated the replacement of methodological hierarchies with the notion of fitness for purpose.

A related aspect of the ‘anti-positivist’ critique is the necessity of judgement, both within scientific research and within the systematic review process itself. As argued by Hammersley (2001), the systematic review methodology has tended to assume that subjectivity is a source of bias that should be minimised by transparency and procedural rules. But the ‘post-positivist’ perspective argues that judgement is unavoidable and need not necessarily be a source of bias. Instead, informed judgement can involve skilled and knowledgeable assessment of what is likely to be true and hence can help to identify mistakes and weed out incorrect and inconsistent assumptions. As a result, both the attempts to minimise judgement within the systematic review process, and the priority given to research evidence as compared to professional knowledge and experience, may be misguided (Hammersley, 2005). As such, this critique questions the rationale of the ‘evidence-based’ process itself.

4.3. Bias in the synthesis of results

Meta-analysis commonly seeks to synthesise the results of various studies by ‘adding’ them together and using the larger effective sample size to obtain more accurate estimates of the size of the relevant effect. This ‘additive’ idea also influences the traditional approach to narrative synthesis. But such an approach is only possible if each of the studies addresses the same specific issue and investigates it in a sufficiently similar way that the findings can be aggregated (Hammersley, 2001). For complex policies and mechanisms operating within heterogeneous social contexts, this may not be viable.

As Hammersley (2001) has argued, the ‘additive’ approach represents a rather impoverished understanding of what synthesis means. Synthesis is more usually understood as combining a set of parts into a whole, which may be greater than the sum of parts. This may involve conceptual development that goes beyond the original studies and is likely to require creative interpretation, rather than simply following a procedure. For example, a study on the effect of energy taxes on household energy use could seek to combine econometric estimates of energy price elasticities, survey data on perceived barriers to technology adoption and behavioural studies of decision-making by individual households. These studies may complement one another, in that they can reinforce or
challenge each other’s assumptions and conclusions (Hammersley, 2005).

Hence, producing a good literature review is not simply a matter of ‘summing’ data, but instead involves judging the validity of different theoretical claims and empirical findings and thinking about how these relate to one another (Hammersley, 2005). An implication of this is that useful syntheses may not require an exhaustive search of the relevant literature, since an attempt to do so may simply lead to diminishing returns. Instead, a more valuable approach may be to use studies in one area to provide insights into problems and issues in another.

5. Is energy different?

Earlier sections have highlighted several differences between energy policy research and those areas where EBPP is most established. These may usefully be summarised as follows:

- **Experimental versus non-experimental**: EBPP has proved most successful in those areas of policy and practice where there is a strong tradition of (quasi-) experimental research. But this is less well represented in energy research, which tends to rely more on econometric analysis of secondary data and economic modelling.

- **Micro versus macro**: EBPP has proved most successful for policies and interventions that are focused at the individual or community level—such as in education and social work. These are also the areas that are most suitable for quasi-experimental research methodologies. But much of energy policy is focused at the sector level and has economy-wide implications.

- **Means versus ends**: EBPP has proved most successful in those areas of policy and practice where there is consensus on a small number of objectives and where the primary topic of debate is the means for achieving those objectives. The contribution of evidence in this context is primarily to establish ‘what works’. But energy policy is characterised by multiple and often competing objectives whose relative importance varies over time and is frequently the topic of fierce debate.

- **Service provider versus regulator**: EBPP has proved most successful in those areas of policy where the government acts as a service provider—such as healthcare and education. But the government’s main role within energy policy is as a regulator and facilitator of competitive markets (Majone, 1997).16

- **Practice versus policy**: EBPP has proved most successful in influencing the practice of doctors, teachers, social workers and other professionals who need to integrate evidence into their everyday decision-making. But with energy policy, politicians, civil servants and regulators focus primarily upon strategic decisions that will shape markets and policy for a considerable period of time.

- **Evaluation versus forecasting**: EBPP has proved most successful in areas where there is a strong tradition of policy evaluation. But energy policy research has placed greater emphasis on economic forecasting than policy evaluation, with the result that the evidence base is weaker.

- **Technical versus politicised**: EBPP has proved most successful in specialised areas of policy or practice, where there is relatively little politicisation—for example, the choice of therapy for a particular disease. But the key questions within energy policy have substantial economic and political consequences and are subject to influence from powerful interest groups. Here, evidence tends to be used to support particular values, interests and policy prescriptions, rather than to underpin decision-making in a more neutral way.

- **Confirmation versus innovation**: EBPP has proved most successful for synthesising evidence from numerous studies that use very similar methodologies. But the funders of energy research increasingly seek empirical, conceptual and methodological innovation, leading to a diversity of approaches that can make the meta-analysis of results problematic.

Each of the above distinctions is an oversimplification and numerous exceptions can be found. For example: systematic reviews have been applied to ‘macro’ policy issues such as inequality and violent crime (Hsieh and Pugh, 1993); to non-service provider areas such as tax policy (Blundell and Walker, 2000); to non-evaluation questions such as the impact of race on sentencing (Sweeney and Haney, 1992); and to highly politicised issues such as GM crops (Levitt, 2003). But while precedents exist in each of these areas, they are often the exception rather than the rule. Similarly, while systematic reviews have been applied to policy areas that have similarities with energy (e.g. transport), the experience to date in these areas is very limited (Kremers et al., 1999).

This suggests that systematic review techniques may need to be modified if they are to be successfully applied to energy policy.

5.1. An example: the rebound effect

The difficulties of applying systematic reviews to energy policy questions may briefly be illustrated with an example. A topic that is attracting increasing attention is the magnitude of the ‘rebound effect’ from energy efficiency improvements. Since improved energy efficiency reduces the effective price of an energy service, there may be an increase in consumption of that service that could offset the energy savings achieved. For consumers, the direct rebound effect may be decomposed into a substitution and income effect, while for producers it may be decomposed into a substitution and an output effect. The issue in

---

16 The author is indebted to Jim Skea, Research Director of the UKERC, for this insight.
question is the size of this effect for a particular energy service, such as household heating or passenger transport.

At first sight, the application of systematic review techniques appears sensible. A large number of empirical studies of the rebound effect is available for both household heating and passenger transport and many of these use econometric or quasi-experimental techniques that appear suitable for appraisal using standard EBPP criteria (Greening et al., 2000). But a closer look reveals that this evidence base is extremely diverse, both empirically and methodologically. The relevant econometric studies use time series, cross-sectional and panel data from a variety of time periods and geographical areas and estimate parameters using a range of functional forms and controlling for a range of intervening variables (Small and Van Dender, 2005). Differences in results owe much to differences in functional specification or empirical application (e.g. time periods before or after oil price shocks), making it difficult to use meta-analysis to estimate the ‘true’ effect (Greening et al., 2000). While these differences could potentially be explored through meta-regression techniques (Espey, 1998), this would need to accommodate both quasi-experimental studies and the larger number of indirect estimates of the rebound effect that are based upon price elasticities. The latter approach rests upon a number of questionable assumptions that may limit both the accuracy and comparability of different studies (Sorrell and Dimitropoulos, 2005). Indeed, the most useful contribution of a review in this area may be to clarify these underlying theoretical issues.

A systematic review would only be possible for the small number of energy services that have been the subject of significant research on rebound effects. Moreover, the precise measurement of the direct rebound effect for a particular energy service may not be the most interesting policy question. This is because the ‘full’ rebound effect also includes:

- **Indirect effects**: The lower effective price of the energy service can lead to changes in the demand for other goods, services and factors of production that also require energy for their provision. For example, the cost savings obtained from a more efficient central heating system may be put towards an overseas holiday.
- **Economy wide effects**: A fall in the real price of energy services may reduce the price of intermediate and final goods throughout the economy, leading to a series of price and quantity adjustments, with energy-intensive goods and sectors gaining at the expense of less energy-intensive ones. Energy efficiency improvements may also reduce energy prices and increase economic growth, which could further increase energy consumption (Greening et al., 2000).

The controversy over the rebound effect relates to the aggregate impact of all these mechanisms. Brookes (2000) and Saunders (2000), for example, argue that improved energy efficiency may actually lead to higher energy consumption overall (‘backfire’).

Systematic review techniques may have little to contribute to this broader, economy-wide question. In addition to the quasi-experimental and price elasticity studies, the relevant evidence base includes: input–output analysis of the energy content of household goods and services (Alfredsson, 2004); decomposition analysis of historical trends in energy consumption (Schipper and Grubb, 2000); general equilibrium modelling of the macro-economy (Grepperud and Rasmussen, 2004); econometric estimates of the elasticity of substitution between energy and capital (Berndt and Wood, 1979); and applications of neoclassical growth theory (Saunders, 2000). Several of these categories have not previously appeared within EBPP, so there is little precedent for comparative evaluation of their methodological quality. Moreover, the overall evidence base is too large in terms of the number of studies, too heterogeneous in terms of approach and empirical application and (most importantly) too embedded in complex and contested theoretical frameworks to be an appropriate subject for a systematic review (Table 6). The central importance of theoretical issues means that a useful review would need to do far more than ‘combine’ different studies to estimate the ‘true’ size of the effect.

This brief review suggests that systematic reviews face considerable methodological difficulties when applied to highly specific questions regarding the direct rebound effect, and are entirely inappropriate for a review of the ‘full’ rebound effect. While the rebound effect may be a particularly difficult topic, it is likely that other policy relevant questions in the energy field have similar characteristics. Similarly, while systematic reviews may be inappropriate for such multidimensional questions, these are also the questions most likely to attract research funding. Hence, the more widespread use of systematic reviews in energy research may require a change in funding priorities.

6. Summary

Energy policy research is not immune from the problems that EBPP set out to address. These include: conflict and confusion over key issues; over-reliance on individual studies; inadequate accumulation and synthesis of research results; and wide-ranging and inconclusive literature reviews that pay insufficient attention to methodological quality (Gross, 2005). To the extent that systematic reviews have a track record of successfully overcoming these

---

17 This makes use of the ‘Khazzoom’ equation: $\eta_E(E) = -\theta p_e(S) - 1$, where $\eta_E(E)$ is the elasticity of energy demand with respect to energy efficiency alone and $\theta p_e(S)$ is the elasticity of energy service demand with respect to the unit price of energy services. Frequently, energy demand ($E$) is substituted for energy service demand $S$, and/or the unit price of energy commodities ($p_e$) is substituted for the unit price of energy services (Sorrell and Dimitropoulos, 2005).
difficulties, there ought to be scope for applying this approach to energy policy questions.

However, systematic reviews have a number of important weaknesses. These include: the narrow range of questions to which they have been applied; the bias towards quantitative research methodologies; the difficulties in addressing complex problems and policies; and the ‘additive’ approach to synthesis that neglects the complementary nature of different studies. Furthermore, there appears to be a mismatch between the type of question for which systematic reviews have been most successful and the type of question that is of greatest interest within energy policy and hence most likely to attract funding.

This suggests that energy researchers may only be able to use systematic review techniques for a subset of questions and may need to modify and extend those techniques when applied. In many cases, the conventional ‘narrative’ review is likely to be more appropriate. The greater use of systematic reviews will require a combination of increased awareness amongst researchers, appropriate training and changes in funding priorities. While systematic reviews may have fundamentally changed the practice of medicine and health-care, there seems little prospect as yet of a comparable impact on energy policy.

Acknowledgements

This research was funded by the UK Research Councils as part of the Technology and Policy Assessment (TPA) function of the UK Energy Research Centre. The author would like to thank Jim Skea (Research Director, UKERC), Rob Gross (Head of TPA, UKERC), Annette Boaz (Queen Mary College), Erik Millstone (SPRU Environment Programme, University of Sussex), Gordon Mackerron, Andy Stirling and John Dimitropoulos (all of the Sussex Energy Group, University of Sussex) for their helpful comments. The usual disclaimers apply.

References


Table 6
An overview of the evidence base for the rebound effect

<table>
<thead>
<tr>
<th>Methodological category</th>
<th>Size of evidence base</th>
<th>Internal diversity</th>
<th>Theoretical content</th>
<th>Suitability for systematic review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation studies</td>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Econometric studies of efficiency and price elasticities</td>
<td>Large</td>
<td>Large</td>
<td>High</td>
<td>Medium/low</td>
</tr>
<tr>
<td>Econometric studies of elasticities of substitution</td>
<td>Large</td>
<td>Large</td>
<td>High</td>
<td>Medium/low</td>
</tr>
<tr>
<td>Input-output analysis</td>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
<td>Low</td>
</tr>
<tr>
<td>General equilibrium modelling</td>
<td>Small</td>
<td>Large</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Neo-classical growth theory</td>
<td>Small</td>
<td>Small</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>

Source: Sorrell and Dimitropoulos (2005).


Petrosino, A., Turpin-Petrosino, C., Finckenauer, J.O., 2000. Well-meaning programmes can have harmful effects! Lessons from experiments of programmes such as Scared Straight. Crime and Delinquency 46 (3), 354–379.


Small, K.A., Van Dender, K., 2005. Study to evaluate the effect of reduced greenhouse gas emissions on vehicle miles traveled. Final Report, ARB Contract Number 02-336, Department Economics, University of Irvine, Irvine, CA.


ANALYSIS

The rebound effect: Microeconomic definitions, limitations and extensions

Steve Sorrell*, John Dimitropoulos

Sussex Energy Group, SPRU (Science & Technology Policy Research), Freeman Centre, University of Sussex, Falmer, Brighton, BN1 9QE, UK

ARTICLE INFO

Article history:
Received 13 June 2006
Received in revised form 16 May 2007
Accepted 19 August 2007
Available online 27 September 2007

Keywords:
Rebound effect
Energy efficiency
Energy demand

ABSTRACT

The rebound effect results in part from an increased consumption of energy services following an improvement in the technical efficiency of delivering those services. This increased consumption offsets the energy savings that may otherwise be achieved. If the rebound effect is sufficiently large it may undermine the rationale for policy measures to encourage energy efficiency.

The nature and magnitude of the rebound effect is the focus of long-running dispute with energy economics. This paper brings together previous theoretical work to provide a rigorous definition of the rebound effect, to clarify key conceptual issues and to highlight the potential consequences of various assumptions for empirical estimates of the effect. The focus is on the direct rebound effect for a single energy service — indirect and economy-wide rebound effects are not discussed.

Beginning with Khazzoom’s original definition of the rebound effect, we expose the limitations of three simplifying assumptions on which this definition is based. First, we argue that capital costs form an important part of the total cost of providing energy services and that empirical studies that estimate rebound effects from variations in energy prices are prone to bias. Second, we argue that energy efficiency should be treated as an endogenous variable and that empirical estimates of the rebound effect may need to apply a simultaneous equation model to capture the joint determination of key variables. Third, we explore the implications of the opportunity costs of time in the production of energy services and highlight the consequences for energy use of improved ‘time efficiency’, the influence of time costs on the rebound effect and the existence of a parallel rebound effect with respect to time. Each of these considerations serves to highlight the difficulties in obtaining reliable estimates of the rebound effect and the different factors that need to be controlled for. We discuss the implications of these findings for econometric studies and argue that several existing studies may overestimate the magnitude of the effect.

© 2007 Elsevier B.V. All rights reserved.

1. Introduction

The rebound effect is the focus of a long-running dispute with energy economics. The question is whether economically worthwhile improvements in the technical efficiency of energy use can be expected to reduce aggregate energy consumption by the amount predicted by simple engineering calculations. For example, will a 20% improvement in the thermal efficiency...
of a heating system lead to a corresponding 20% reduction in energy consumption? Economic theory suggests that it will not. Three separate mechanisms may reduce the aggregate energy savings achieved (Greening et al., 2000):

- **Direct rebound effects**: Improved energy efficiency for a particular energy service will decrease the effective price of that service and should therefore lead to an increase in consumption of that service. This will tend to offset the reduction in energy consumption provided by the efficiency improvement.

- **Indirect effects**: The lower effective price of the energy service may lead to changes in the demand for other goods, services and factors of production that also require energy for their provision. For example, the cost savings obtained from a more efficient central heating system may be put towards an overseas holiday.

- **Economy wide effects**: A fall in the real price of energy services may reduce the price of intermediate and final goods throughout the economy, leading to a series of price and quantity adjustments, with energy-intensive goods and sectors likely to gain at the expense of less energy-intensive ones.

Numerous empirical studies, principally from the US, suggest that these rebound effects are real and can be significant (Greening et al., 2000). However, while their basic mechanisms are widely accepted, their magnitude and importance are disputed. Some analysts argue that rebound effects are of minor importance for most energy services (Schipper and Grubb, 2000), while others argue that the economy-wide effects can be sufficiently important to completely offset the energy savings from improved energy efficiency (Brookes, 1990; Saunders, 1992). The policy implication is that non-price regulations to improve energy efficiency may neither reduce energy demand nor help to mitigate climate change.

Indirect and economy-wide rebound effects involve general equilibrium adjustments that are difficult to analyse empirically. In contrast, direct rebound effects can be investigated more directly through quasi-experimental studies or the econometric analysis of secondary data. However, such studies raise a number of definitional and methodological issues that are inadequately discussed in the literature. The disagreement over the magnitude and importance of the rebound effect may result in part from lack of clarity over these basic definitions and issues. Moreover, since many empirical studies overlook key methodological issues, their estimates of the rebound effect could potentially be biased.

This paper examines the definition and measurement of the direct rebound effect for individual energy services. Indirect and economy-wide effects are not discussed. The focus throughout is on energy efficiency improvements in consumer goods such as cars and central heating systems, since this is where the bulk of the empirical evidence lies. While analogous arguments apply to energy efficiency improvements by producers, the evidence here is weaker and harder to interpret (Greening and Greene, 1998).

The paper is structured as follows. First, we present a general ‘household production’ framework for characterising the demand for energy services that helps to illustrate the different trade-offs involved. Second, we show how the direct rebound effect can be represented as an efficiency elasticity of energy demand and how it may be decomposed into the sum of elasticities for the number, capacity and utilisation of energy conversion devices. Third, we show the relationship between the rebound effect and the price elasticity of the demand for ‘useful work’, as well as the price elasticity of the demand for energy, and show why empirical studies using these definitions provide a primary source of evidence for the direct rebound effect. We then expose the limitations of these definitions, focusing on: a) the potential correlation between various input costs and improvements in energy efficiency; b) the endogeneity of energy efficiency and the implied need for simultaneous equation estimation; and c) the role of time costs and time efficiency in the production and consumption of energy services. We identify some of the factors that need to be controlled for to obtain accurate estimates of the rebound effect and argue that the neglect of these factors by several existing studies may lead the rebound effect to be overestimated.

2. The demand for energy services

The demand for energy (E) derives from the demand for energy services (ES) such as thermal comfort, refrigeration and motive power. These services, in turn, are delivered through a combination of energy commodities and the associated energy systems, including energy conversion devices. Consumers are assumed to derive utility from consuming these services, rather than from consuming energy commodities and other market goods directly. In practice, nearly all services require energy in some form, although energy may form a much smaller proportion of total costs for some services than for others.

An essential feature of an energy service is the useful work (S) obtained, which may be measured by a variety of thermodynamic or physical indicators (Patterson, 1996). These indicators may, in turn, be decomposed in a variety of ways to reveal the relative importance of different contributory variables. For example, the useful work from the private cars owned by a group of households may be:

- Measured in vehicle kilometres and decomposed into the product of the number of cars (NO) and the mean driving distance per car per year (UTIL): $S = NO \times UTIL$
- Measured in passenger kilometres and decomposed into the product of the number of cars (NO), the mean driving distance per car per year (UTIL) and the average number of passengers carried per car (LF): $S = NO \times UTIL \times LF$
- Measured (rather unconventionally) in tonne kilometres and decomposed into the product of the number of cars (NO), the mean driving distance per car per year (UTIL) and the mean (loaded or unloaded) vehicle weight (CAP): $S = NO \times CAP \times UTIL$

---

3 Indirect effects may be addressed through the use of input-output models (Kok et al., 2006), while economy-wide effects may be addressed through the use of Computable General Equilibrium (CGE) models (Grepperud and Rasmussen, 2004). In both cases, the existing evidence base is very small. For reviews of input-output approaches to rebound effects see Sorrell and Dimitropoulos (2007), while for CGE approaches see Allan et al. (2007).
In practice, the choice of indicator and associated decomposition will depend upon the objective of the analysis, the level of aggregation (e.g. household, sector, economy) and the availability of the relevant data. In much empirical work, measures of useful work are not decomposed.

It is important to recognize that energy services also have broader attributes (A) that may be combined with useful work in a variety of ways. For example, all cars deliver passenger kilometres, but they may vary widely in terms of features such as speed, comfort, acceleration and prestige. The combination of useful work (S) with these associated attributes (A) provides the full energy service: ES=es(S,A).

Becker’s ‘household production’ model provides a useful framework for understanding the demand for energy services (Becker, 1965). In this model, which is briefly outlined in Annex A, individual households are assumed to produce useful work by combining energy, capital and other market goods, together with some of the household’s own time. For example, mobility may be produced by the household through the combination of gasoline, a private car, expenditure on maintenance and driving time. Similarly, a cooked meal may be produced through the combination of natural gas, a gas cooker, ingredients and cooking time. The provision of useful work for a particular energy service may then be described by a production function, representing the maximum output that can be obtained from the currently available technology for a given level of energy and other inputs (Wirl, 1997). But the provision of broader attributes for a given amount of useful work is likely to require additional inputs; or, alternatively, for a given budget, the provision of broader attributes is likely to reduce the amount of useful work.

The primary contribution of this model in the present context is to emphasise that consumption of an energy service involves three interrelated trade-offs, namely:

- Between consumption of useful work versus consumption of other attributes of an energy service;
- Between energy, capital, other market goods and time into the production of an energy service; and
- Between consumption of different types of energy service.

This general framework forms the foundation for what follows.

3. The rebound effect as an efficiency elasticity

The energy efficiency (ɛ) of an energy system may be defined as $\eta = S/E$, where E represents the energy input required for a unit output of useful work (however measured).\(^5\) For example, a car may require 10 litres of gasoline to drive 100 km. The energy cost of useful work ($P_E$) is then given by $P_E = P_S/ɛ$, where $P_S$ represents the price of energy. This is one component of the total cost of useful work, which also includes other input costs, such as annualised capital costs, maintenance costs and time costs.

Consider a situation where the energy efficiency of an energy system is improved ($\Delta \epsilon > 0$), but the costs of non-energy inputs and the consumption of other attributes of the energy service remain unchanged. In the absence of a rebound effect, the demand for useful work would remain unchanged ($\Delta S = 0$) and energy demand would be reduced in proportion to the improvement in energy efficiency ($\Delta E/E = -\Delta \epsilon/\epsilon$). But the efficiency improvement lowers the energy cost per unit of useful work ($\Delta P_S < 0$) and hence also the total cost. Assuming that the energy service has a price elasticity in the normal range, consumers will demand more useful work ($\Delta S > 0$) and the proportional change in energy consumption will be less than the proportional change in energy efficiency ($\Delta E/E < -\Delta \epsilon/\epsilon$).

The change in demand for useful work following a small change in energy efficiency may be measured by the efficiency elasticity of the demand for useful work ($\eta_S(S)$):

$$\eta_S(S) = \frac{\partial S/E}{\partial \epsilon}$$

In a similar manner, the change in energy demand following a small change in energy efficiency may be measured by the efficiency elasticity of the demand for energy ($\eta_E(E)$):

$$\eta_E(E) = \frac{\partial E/S}{\partial \epsilon}$$

Substituting $E = S/\epsilon$ in the equation for $\eta_E(E)$ and taking partial derivatives we can derive the following relationship between these two elasticities: \(^3\)

Definition 1.

$$\eta_E(E) = \eta_S(S) - 1.$$ 

The efficiency elasticity of the demand for useful work ($\eta_S(S)$) has been commonly taken as a direct measure of the rebound effect (Berkhout et al., 2000). The actual saving in energy consumption will only be equal to the predicted saving from engineering calculations when this elasticity is zero ($\eta_S(S) = 0$). Under these circumstances, the efficiency elasticity of the demand for energy ($\eta_E(E)$) is equal to minus one. A positive rebound effect implies that $\eta_S(S) > 0$ and $\eta_E(E) < 1$. For example, a positive rebound effect for car travel implies that improvements in vehicle fuel efficiency increase the demand for vehicle kilometres, with the result that the savings in energy consumption are less than predicted from engineering calculations alone.

If the demand for the energy service is inelastic ($0 < \eta_S(S) < 1$) improvements in energy efficiency should reduce energy demand ($0 > \eta_E(E) > -1$). But if the demand for the energy service is elastic ($\eta_S(S) > 1$), improvements in energy efficiency will actually increase energy consumption. This somewhat counter-intuitive outcome is termed ‘backfire’ in the literature (Saunders, 1992).

Technological improvements in energy efficiency may lead to an increase in the number of energy conversion devices (NO), their average size (CAP), their average utilisation (UTIL) and/or their average load factor (LF). For example, people may buy more

\(^5\) The appropriate measure of energy efficiency depends upon the objectives of the analysis and is generally a property of the energy system, rather than just the energy conversion device. For example, if the average internal temperature is taken as the appropriate measure of useful work from a household heating system, energy efficiency will depend upon both the thermal efficiency of the boiler and the level of thermal insulation.

\(^3\) See Annex B for derivations of this and subsequent definitions and formulae.
and hence the demand (paribus assumptions given above, the effect on the total cost of energy efficiency is constant. Under the ceteris-paribus assumptions (e.g. more SUVs) as well as decreases in average load factor (e.g. less car sharing) will be overlooked.\footnote{The first of these rebound effects could be captured if useful work for private travel was measured in unloaded tonne kilometres rather than vehicle kilometres. This would be possible if data was available on the composition of the vehicle stock and the average unloaded weight of different types of vehicle. The second effect could be captured if useful work was measured in passenger kilometres rather than vehicle kilometres. This would require data on the average load factor of different types of vehicle. To capture both of these rebound effects, useful work would need to be measured in loaded tonne kilometres.}

\begin{equation}
\eta_\varepsilon(E) = \left[ \eta_\varepsilon(NO) + \eta_\varepsilon(CAP) + \eta_\varepsilon(UTIL) \right] - 1.
\end{equation}

The relative importance of these variables may vary widely between different energy services and over time. For example, technological improvements in the energy efficiency of new refrigerators are unlikely to increase the average utilisation of the refrigerator stock (measured in hours/year) but could lead to an increase in both the number and average size of refrigerators over time (since the cost per cubic metre of refrigeration has reduced). The majority of empirical estimates of the rebound effect relate to travel by private cars, where useful work is commonly measured in vehicle kilometres travelled and decomposed into the product of vehicle numbers and the mean distance travelled per car per year (Greene et al., 1999b; Small and Van Dender, 2005). An important consequence of this is that any increases in average vehicle weight as a result of energy efficiency improvements (e.g. more SUVs) as well as decreases in average load factor (e.g. less car sharing) will be overlooked.\footnote{The first of these rebound effects could be captured if useful work for private travel was measured in unloaded tonne kilometres rather than vehicle kilometres. This would be possible if data was available on the composition of the vehicle stock and the average unloaded weight of different types of vehicle. The second effect could be captured if useful work was measured in passenger kilometres rather than vehicle kilometres. This would require data on the average load factor of different types of vehicle. To capture both of these rebound effects, useful work would need to be measured in loaded tonne kilometres.}

The marginal utility of energy service consumption is likely to decline with increased consumption, which should reduce the direct rebound effect from energy efficiency improvements. For example, rebound effects from improvements in the energy efficiency of household heating systems should decline rapidly once whole-house indoor temperatures approach the maximum level for thermal comfort. One implication, frequently observed in the policy evaluation literature, is that direct rebound effects will be higher among low income groups, since these are further from saturation in their consumption of individual energy services (Boardman and Milne, 2000).

### 4. The rebound effect as a price elasticity

Since \( P_S = P_C / \varepsilon \), raising (lowering) energy efficiency (\( \varepsilon \)) when energy prices (\( P_S \)) are constant should have the same effect on the energy cost of useful work (\( P_S \)) as falling (rising) energy prices when energy efficiency is constant. Under the ceteris-paribus assumptions given above, the effect on the total cost and hence the demand (\( S \)) for useful work should be symmetrical. If other inputs are held constant, we can write the demand for useful work solely as a function of energy prices and energy efficiency: \( S = f(P_S / \varepsilon) \). The demand for energy is then given by: \( E = g(P_S / \varepsilon) \). Assuming that energy prices are exogenous (i.e. \( P_S \) does not depend upon \( \varepsilon \)), we can differentiate this equation with respect to energy efficiency to give an alternative definition of the rebound effect:

\begin{equation}
\eta_\varepsilon(E) = -\eta_{PE}(S) - 1.
\end{equation}

Hence, under these assumptions, the efficiency elasticity of energy demand (\( \eta_\varepsilon(E) \)) is equal to the energy cost elasticity of the demand for useful work (\( \eta_{PE}(S) \)), minus one. Effectively, the negative of the energy cost elasticity for useful work (\( \eta_{PE}(S) \)) being used as a proxy for the efficiency elasticity of useful work (\( \eta_\varepsilon(S) \)), which in turn is the primary definition of the rebound effect. If useful work is a normal good, we expect that \( \eta_{PE}(S) \leq 0 \).

For example, if the elasticity of vehicle km (\( S \)) with respect to energy cost per kilometre (\( P_S \)) is estimated as \(-0.10\), then the elasticity of gasoline demand with respect to fuel efficiency can be estimated from Definition 3 as \(-0.90\). This implies that the demand for gasoline will fall by only 9% if the fuel efficiency of vehicles improves by 10% — or, alternatively, that 10% of the potential savings in gasoline consumption will be ‘taken back’ by increased vehicle use.

A version of this expression is derived by Khazzoom (1980), Berkhout et al. (2000), Binns wander (2003) and Greene et al. (1999a) and is generally used in preference to Definition 1 in empirical estimates of the rebound effect. For many energy services, the available data provides only limited variation in the independent variable for Definition 1 (\( \varepsilon \)) while at the same time requiring energy prices to be controlled for. In contrast, the data provides much greater variation in the independent variable for Definition 3 (\( P_S \)) since this reflects both variations in energy efficiency and variations in energy prices. For many energy services, the historical and cross-sectional variations in the relevant energy commodity prices tend to be much greater than the corresponding variations in the energy efficiency of the relevant energy systems. Given the assumption that consumers respond in the same way to increases (decreases) in energy prices as to decreases (increases) in energy efficiency, Definition 3 provides a means to estimate the potential magnitude of rebound effects from efficiency improvements even in circumstances where the available data provides little or no variation in energy efficiency.

Empirical studies based upon Definition 3 require accurate measures of both the demand for useful work (\( S \)) for the relevant energy service and the energy cost per unit of useful work (\( P_S \)). The latter, in turn, depends upon energy commodity prices and the energy efficiency of the relevant energy system. But, depending upon how it is defined, the measurement of useful work for many types of energy service can be problematic. For example, the useful work from a domestic heating system could be defined as the average internal temperature of the house and measured directly using field thermometers or indirectly from thermostat settings. But the latter are notoriously inaccurate and can be a poor proxy for the thermal comfort of the occupants, which depends upon other variables such as humidity and airflow (Greening and Greene, 1998). One reason that travel by private car (in the United States) is the most widely studied
area for the rebound effect is that relatively good data is available on vehicle kilometres as a measure of useful work, while fuel cost per kilometre is easily estimated by combining data on gasoline prices and vehicle fuel efficiency (Greene, 1992).

While obtaining measures of useful work (S) can be difficult, data is more commonly available on the energy demand (E) for the relevant energy service. For example, data may be available on the demand for gas for household heating (although the use of gas for cooking could provide a complication). If we assume that energy efficiency is constant, the symmetry argument implied by the ratio \( \frac{P_S}{E} = \frac{P_F}{F} \) leads to an alternative definition for the rebound effect based upon the own price elasticity of energy demand:

**Definition 4.**

\[
\eta_s(E) = -\eta_P(E) - 1.
\]

It is this expression, rather than Definition 2, that was originally put forward by Khazzoom and is also used by Wirl (1997) in his comprehensive analysis of the economics of energy efficiency. Definition 4 shows that under certain assumptions, the rebound effect may be approximated by the own price elasticity of energy demand for the relevant energy service. Note that this definition is only meaningful when the energy demand in question relates to a single energy service (e.g. refrigeration). In practice, available measures of energy demand frequently apply to a collection of energy services (e.g. household electricity use), although techniques such as conditional demand analysis may allow the proportion of demand attributable to an individual service to be estimated (Parti and Parti, 1980).^5^

Empirical studies of the rebound effect for different energy services may use either Definitions 1, 3 or 4, but the differences between studies are not always made clear. For example, in their comprehensive literature review of the rebound effect, Greening and Greene (1998) cite 23 studies of household heating, but place particular weight on the methodologically rigorous studies of household survey data by Klein (1987), Hseuh and Gerner (1993) and Schwarz and Taylor (1995). As Table 1 makes clear, these three studies use different definitions of the dependent and independent variable, apply different methodologies and controls and focus upon different fuels. Only one of the three studies (Schwarz and Taylor) represents an explicit investigation of rebound effects (the other two do not mention the term) and Greene and Greening’s estimate of rebound effects is based upon a different definition in each case. Clearly, the existing literature is too small and diverse to allow a consistent approach to this problem.

**Table 1 – Varying estimates of the rebound effect for household heating**

<table>
<thead>
<tr>
<th>Study</th>
<th>Dependent variable</th>
<th>Approach</th>
<th>Greening and Greene (1998) estimate from:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Klein (1987)^a</td>
<td>Proxy measure for S from thermostat setting</td>
<td>Simultaneous estimation of a cost function for S, a demand function for S and an equation for the relative share of capital and fuel</td>
<td>( \eta_S(S) )</td>
</tr>
<tr>
<td>Hseuh and Gerner (1993)^b</td>
<td>E estimated from energy bill (No measure of S available)</td>
<td>Estimates a reduced form equation for the demand for E. Incorporating engineering and other variables that affect the demand for S.</td>
<td>( \eta_E(E) )</td>
</tr>
<tr>
<td>Schwarz and Taylor (1995)^c</td>
<td>Proxy measure for S from thermostat setting</td>
<td>Estimate an equation for the demand for S, that incorporates a variable representing the thermal resistance of the house</td>
<td>( \eta_S(S) )</td>
</tr>
</tbody>
</table>

^a Space heating demand is defined as the difference between the internal thermostat setting and the external temperature during the heating season. Simultaneous equations are required because S appears in both the cost function and factor share equation and individual household attributes affect both the production and consumption decision. The study estimates the elasticity of demand for useful work with respect to the total cost of useful work, but Greening and Greene (1998) interpret this incorrectly as \( \eta_S(S) \). The study does not mention the rebound effect directly.

^b In principle, the reduced form equation should allow \( \eta_S(S) \) to be estimated — where \( e \) represents the overall energy efficiency of the household, including both building fabric and energy conversion devices. But Hseuh and Gerner only quote the change in energy consumption following a physical change in one element of the system, such as increasing insulation thickness by one inch. So instead of \( \eta_S(E) \), Greene and Greening take the estimate of \( \eta_E(E) \) as a measure of the rebound effect for electricity only, while the estimate for natural gas is ignored.

^c The measure of energy efficiency is the thermal resistance of the house. The specification allows the efficiency elasticity of the thermostat setting and the efficiency elasticity of demand for useful work to be estimated. The difference between each of these and \( \eta_E(E) \) are taken as alternative measures of the rebound effect.

---

\[ \text{For example, Haas et al., (1998) unbundled energy use for space heating from that for water heating by assuming that the latter was constant over the year, while the former depended upon external temperature.} \]
Gately, 1994, 1995; Haas and Schipper, 1998). For example, Dargay (1992) found that the reduction in UK energy demand following the price rises of the late 1970s was five times greater than the increase in demand following the price collapse of the mid-1980s. Two explanations for this are that higher energy prices induce technological improvements in energy efficiency that are not reversed when energy prices fall, while investment in measures such as thermal insulation is largely irreversible over the short to medium-term (Grubb, 1995). Energy efficiency requirements may also become embodied in regulations that ensure that new investments maintain high standards, even in the absence of a price incentive. As a result, estimates of price elasticities based upon time series data are likely to vary according to whether energy prices were rising, falling (or both) over the period in question (Haas and Schipper, 1998). In the case of the rebound effect, the appropriate proxy for energy efficiency improvements is reductions in energy prices. Hence, empirical estimates based upon periods of rising energy prices are likely to overestimate the size of the effect.

Econometric estimates based upon Definitions 3 and 4 are a primary source of evidence for the rebound effect. Hence, the assumptions behind these definitions – and particularly the argument of symmetry between changes in energy efficiency and changes in energy prices – require careful scrutiny. The following three sections explore the limitations of these definitions in more detail, focusing on:

- The correlation between energy efficiency and other input costs, notably capital costs;
- The endogeneity of energy efficiency and the implied need for simultaneous equation estimation; and
- The role of time costs and time efficiency in the production and consumption of energy services.

5. Correlation between energy efficiency and other input costs

For an individual energy service, changes in energy commodity prices are unlikely to be correlated with changes in other input costs or with changes in the broader attributes of the energy service. But the same cannot be said about changes in energy efficiency. In many (although by no means all) cases, energy efficiency conversion devices will have a higher capital cost than inefficient models (i.e. c and k will be positively correlated). For example, UK building regulations now require high efficiency condensing boilers to be used when installing or replacing a domestic central heating system and these have historically cost some £200-300 more than a conventional boiler.

Khazzoom (1980) assumed this problem away by arguing that a more energy efficient appliance does not necessarily entail a greater initial cost and citing the lower cost of smaller and more fuel-efficient cars as an example. But in this case, the improvement in energy efficiency is likely to have been achieved at the expense of other attributes of the energy service, such as carrying capacity and legroom (i.e. e and a will be negatively correlated). In general, improvements in energy efficiency could result from energy-saving technological change, substitution between energy and other inputs, or substitution between useful work and other output attributes. In practice, many energy services have multiple attributes (e.g. size, comfort, reliability, speed) and each attribute may have a non-zero elasticity with respect to the energy cost of useful work. As Einhorn (1982) has argued, the long-term response to a reduction in energy costs will depend upon the trade-offs between useful work and these multiple attributes.

Khazzoom’s neglect of capital costs has been challenged by several authors (Besen and Johnson, 1982; Einhorn, 1982; Henly et al., 1988; Lovins et al., 1988) who argue that it may lead empirical studies that rely upon Definitions 3 and 4 to overestimate the rebound effect. Henly et al. (1988) illustrate this clearly by including annualised capital costs (Pa) in the equation for energy demand. Assuming that capital costs are a function of energy efficiency, the basic identity becomes: E = E[Pa/E, Pa] / E. We can then derive the following alternative definition of the efficiency elasticity of energy demand:

**Definition 5.**

\[ \eta_t(E) = -1 - \eta_{g}(S) - \left[ \eta_{g}(S) \eta_{t}(Pa) \right] \]

Compared to Definition 3 (and Definition 4), there is an additional term in square brackets. This is the product of the elasticity of demand for useful work with respect to capital costs (\( \eta_{t}(S) \)) and the elasticity of capital costs with respect to energy efficiency (\( \eta_{t}(Pa) \)). We expect the first of these to be negative: higher capital costs should reduce the long-run demand for useful work, largely because they should reduce the number of energy conversion devices (\( \eta_{t}(S) \)) and/or their average size (\( \eta_{t}(CAP) \)). Under the assumption that energy efficient equipment is more expensive, the second term will be positive, making the product of these two expressions negative. The net result will be to reduce the absolute magnitude (\( \eta_{t}(E) \)) of the efficiency elasticity of energy demand. Hence, if energy efficient equipment is more expensive, the rebound effect may be smaller than suggested by studies that rely primarily upon historical or cross-sectional variations in energy prices and estimate the rebound effect from Definitions 3 and 4. This is because the change in energy service demand following a change in energy prices will be different from that following a change in energy efficiency. In general, such studies will tend to overestimate the rebound effect. The size of this upward bias will depend upon the relative magnitude of the three separate elasticities.

The correlation between energy efficiency and capital costs may be expected to vary between energy services and over time. In areas such as computing, for example, improvements in energy efficiency have long been associated with both improvements in service attributes and reductions in capital costs (Triplett, 1989). Also, higher capital costs will only reduce the rebound effect if the consumer faces the full cost of the purchase decision. If, for example, the additional cost of energy efficient conversion devices is fully subsidised, the higher initial cost...
should not affect the purchase decision. Furthermore, if government subsidies make energy-efficient devices cheaper than inefficient models, it is possible that the rebound effect will be amplified (i.e. if both $\eta_p(S)$ and $\eta_p(P_2)$ are negative, their product will be positive). Empirical support for this is arguably provided by Roy’s (2000) study of rebound effects for rural lighting in India.

Consideration of the role of capital costs further highlights the importance of distinguishing between the number, capacity and utilisation of energy service devices when estimating rebound effects. Once an appliance is purchased, the capital cost is sunk and hence should be irrelevant to the utilisation decision. But higher capital costs may lead to the purchase of fewer, smaller and/or different conversion devices, depending upon the trade-offs between different categories of input costs and between useful work and other output attributes. Holding output attributes constant, energy efficient conversion devices allow their owners to enjoy a greater consumer surplus in each time period, owing to the higher demand for useful work (Einhorn, 1982). But if the more efficient appliance is also more expensive than the inefficient alternative, it will only be purchased by an ‘economically rational’ consumer if the present value of the discounted stream of additional consumer surplus exceeds the present value of the additional capital cost. A mandatory requirement for new capital equipment to meet high standards of energy efficiency could mean that consumers will choose to delay replacing their existing equipment (if owned and if still working); choose to purchase smaller or different equipment; choose to purchase inefficient, second-hand equipment; or choose to go without the energy service altogether. The net effect on energy consumption for the relevant energy service could therefore be ambiguous. But in all cases, the effect of the efficiency standard will be different from that of a change in energy prices, so an estimate of the rebound effect based upon the latter is likely to be incorrect.

It is also possible that improvements in energy efficiency will be associated with changes in other input costs, such as operation and maintenance (O&M) costs. If, for example, more efficient conversion devices are less reliable and more costly to maintain and operate, the effect of efficiency improvements on the demand for useful work will again be different from the effect of changes in energy prices. However, the evidence for a positive correlation between energy efficiency and O&M costs is absent for most energy services, and for some the correlation may be negative. In general, the magnitude and direction of the bias in estimating the rebound effect using Definitions 3 and 4 will depend upon the degree and sign of the correlation between energy efficiency and all other categories of input costs. If they are positively correlated, the bias will be negative and the rebound effect will be underestimated, while if they are negatively correlated the bias will be positive and the rebound effect underestimated.

Even if improvements in energy efficiency are not associated with changes in other input costs, certain types of rebound effect may be constrained by the real or opportunity costs associated with increasing the demand for useful work. Two important examples are the opportunity cost of space (e.g. increasing refrigerator size may not be the best use of available space) and the opportunity cost of time (e.g. driving longer distances may not be the best use of available time). Both of these examples point to physical constraints on the demand for certain categories of useful work by individual households. However, space constraints may become less important over time if technological improvements reduce the average size of conversion devices per unit of useful work (e.g. computing) or if rising incomes lead to an increase in average living space (e.g. compare refrigerator sizes in the US and the UK (Wilson and Boehland, 2005). In contrast, while technological improvements may reduce the time requirements per unit of useful work, the opportunity cost of time should increase with rising incomes. The relationship between time constraints and energy service consumption appears particularly important and is discussed further below.

6. Endogenous energy efficiency

Definitions 1 and 3 assume that energy efficiency is independent of the values of other independent variables — in other words, that it is exogenous. This follows naturally from Khazoom’s original focus on the effect of mandatory energy efficiency standards for household appliances. In practice, however, the level of energy efficiency is likely to be influenced by one or more of the other dependent variables — in other words, energy efficiency must be considered partly endogenous. In particular, energy efficiency may be expected to be a function of current and historical energy prices: $\epsilon(P_2)$ (Greene et al., 1999b; Small and Van Dender, 2005).7

If the demand for useful work depends upon the energy cost of useful work ($S=s(P_2/\epsilon)$) and energy efficiency depends upon energy prices ($\epsilon=s(P_2/\epsilon)$), the demand for energy for the relevant energy service may be represented as $E=S/\epsilon=s(P_2/\epsilon)/\epsilon(P_2)$. If we differentiate this expression with respect to energy prices and substitute the resulting expression for $\eta_p(S)$ into Definition 3, we obtain an alternative definition of the rebound effect that takes into account price-induced energy efficiency improvements:

Definition 6.

$$\eta_p(E) = - \left[ \eta_p(E) + \eta_p(\epsilon) \frac{1}{1 - \eta_p(\epsilon)} \right] - 1.$$  

Where the expression in square brackets represent the energy cost elasticity of the demand for useful work ($\eta_p(S)$).

Previous versions of this equation have appeared in Blair et al. (1984), Mayo and Mathis (1988) and Small and Van Dender (2005). In principle, Definition 6 provides an alternative method of estimating the rebound effect. Rather than estimating the energy cost elasticity of the demand for useful work, one could separately estimate the own price elasticity of energy consumption for the relevant energy service ($\eta_p(E)$) and the elasticity of energy efficiency with respect to energy prices

7 In the short term, increases in energy commodity prices may encourage consumers to utilise existing equipment in more energy efficient ways — such as increasing average load factor (e.g. car sharing), or adopting energy efficient operating practices (e.g. avoiding excessive speed). In the longer term, consumers may choose to purchase more energy efficient conversion devices, while producers may choose to devote expenditure to developing, improving and marketing such devices.
The resulting calculated value for the energy cost elasticity of the demand for useful work \((\eta_\ell(S))\) could then be used to estimate the rebound effect.

It is clear from Definition 6 that the energy cost elasticity of the demand for useful work \((\eta_\ell(S))\) will only be equal to the own price elasticity of the demand for energy for the relevant energy service \((\eta_\ell(E))\) if the energy price elasticity of energy efficiency is equal to zero \((\eta_\ell(e)=0)\). This is unlikely to be the case in practice. Hanley et al. (2002) have derived an expression for the relative magnitude of different price elasticities that should hold for all econometric estimates:

\[
\left| \eta_\ell(S) \right| \leq \left| \eta_\ell(S) \right| \leq \left| \eta_\ell(E) \right| \leq \left| \eta_\ell(E) \right|.
\]  

(3)

This relationship provides a useful point of reference for the results from individual studies and is supported by evidence from recent surveys (Espey, 1998; Hanley et al., 2002; Graham and Glaister, 2004). It suggests that the elasticity of the demand for useful work with respect to energy costs should be smaller than the elasticity of energy demand with respect to energy prices. This shows that, relative to Definition 3, Definition 4 is likely to overestimate the magnitude of the rebound effect due to the neglect of price-induced energy efficiency improvements.

It seems likely that energy efficiency will also be a function of other endogenous or exogenous variables in ways that could bias the results of empirical studies (Small and Van Dender, 2005). In particular, if consumers expect to have a high demand for useful work, they may be more likely to choose an energy-efficient conversion device in order to minimise the energy cost of useful work. For example, drivers may choose to purchase a more fuel-efficient car if they expect to drive long distances.\(^8\) This may create a positive correlation between \(S\) and \(e\) that is in addition to the positive correlation created by the direct rebound effect. If this is not corrected for in empirical studies, the magnitude of the rebound effect will again be overestimated. Moreover, as pointed out by Small and Van Dender (2005), this type of endogeneity makes the logic behind Definition 3 circular: the demand for useful work \((S)\) depends upon the energy cost of useful work \((P_\ell)\), which in turn depends upon energy efficiency \((e)\) which in turn depends upon the demand for useful work \((S)\).

This simultaneous determination of an endogenous variable \((e)\) with another endogenous variable \((S)\) can be captured with a simultaneous equation model. This starts with a set of \(n\) equations for \(n\) endogenous variables, with each equation representing either a causal relationship or an equilibrium condition. Such models could be formulated in a variety of ways, depending upon data availability. Small and Van Dender (2005), for example, established separate equations for the number \((\text{NO})\) of private cars, their total annual mileage \((S)\) and the average fuel efficiency of the car fleet \((e)\) (changes in vehicle size were ignored). They base their model upon the following generic assumptions regarding consumer choices:

- The total demand for useful work \((S)\) is influenced by the number of energy conversion devices \((\text{NO})\), the energy cost of useful work \((P_\ell=P_\ell/e)\) and a number of exogenous variables \((X_\ell)\).
- The number of energy conversion devices \((\text{NO})\) is influenced by the capital cost of those devices \((P_\ell)\), the anticipated demand for useful work \((S)\), the energy cost of useful work \((P_\ell/e)\) and a number of exogenous variables \((X_\ell)\).
- The average efficiency of the stock of conversion devices \((e)\) is influenced by the price of energy \((P_\ell)\), the anticipated demand for useful work \((S)\), regulatory standards on the energy efficiency of new devices \((R_\ell)\) and a number of exogenous variables \((X_\ell)\).

This leads to the following set of ‘structural’ equations:

\[
S = s(\text{NO}, (P_\ell/e), X_\ell) \\
\text{NO} = no( P_\ell, S, (P_\ell/e), X_{\text{NO}}) .
\]  

(4)

It is an empirical question as to whether a simultaneous equation model is appropriate for a particular energy service. In some cases, the joint dependence of some or all of the variables may either not hold or be sufficiently weak that it can be ignored. For example, Johansson and Schipper (1997) assumed that mean driving distance per vehicle was a function of the number of vehicles and their average fuel efficiency, but argued that the latter did not depend upon mean driving distance because: ‘...one chooses what distance to drive for a given vehicle stock with different characteristics, and not the other way round’ (Johansson and Schipper, 1997). In contrast, Small and Van Dender (2005), Greene et al. (1999b) and Wheaton (1982) all formulate models in which energy efficiency is a function of the number of cars and distance driven and each find the relevant coefficients to be statistically significant.\(^9\)

The key point, however, is that if joint dependence is relevant, the equations need to be estimated through an appropriate instrumental variable technique. If, instead, one or more of the individual equations are estimated through ordinary least squares (OLS), the resulting coefficients will be biased and inconsistent owing to serial correlation between a regressor and the residuals. As an illustration, Small and Van Dender (2005) found that the use of OLS in their model overestimated the short and long-run rebound effects for car travel by 88% and 53% respectively (although factors other then endogeneity may have been involved).

The use of a simultaneous equation model provides a clearer understanding of the implications of changes in energy efficiency, whether induced by regulatory intervention, energy price increases or other factors. For example, a mandatory standard for the energy efficiency of new conversion devices will have a direct effect on the energy efficiency of the stock, through the third of the equations in (4). However, improvements in energy efficiency will also tend to increase the number of conversion devices, which in turn will increase the total demand for useful work. Improvements in energy efficiency should also increase the demand for useful work by reducing the associated energy costs. The net increase in the demand for useful work will in turn encourage higher energy efficiency.

\(^8\) This is a hypothesis to be tested. A counter argument could be that drivers will purchase larger cars if they expect to drive long distances, since these are more comfortable. As larger cars tend to be less fuel-efficient, this may lead to a negative correlation between \(S\) and \(e\).

\(^9\) However, Small and Van Dender find no support for the endogeneity implied by the second of the equations in (9), since the coefficients on \(P_\ell\) and \(S\) are not significant.
Hence, a change in an exogenous variable such as regulatory standards for energy efficiency may trigger a complex set of changes within the system until a new equilibrium is reached. If the behavioural assumptions given above hold, the total change in energy efficiency following the regulatory intervention will be greater than the direct change, as will the total change in energy service demand.

The structural equations may be solved to allow each of the endogenous variables to be written solely as functions of the exogenous variables, giving so-called ‘reduced form’ equations. However, many empirical estimates of the rebound effect use neither a structural equation system nor their reduced form solution. Instead, they employ what Small and Van Dender (2005) term a ‘partially reduced form’ equation for \( S \), denoted here by the symbol \( \tilde{s} \). This includes energy efficiency indirectly via the energy cost of useful work, but does not include the number of conversion devices:

\[
S = \tilde{s}(P_E, \frac{P_T}{\varepsilon}, X_{NO}, X_S).
\]

Since energy efficiency is endogenous, estimation of this equation by OLS is likely to lead to biased estimates of the rebound effect. Moreover, the bias will be compounded if (as is commonly the case), capital costs \( (P_W) \) or other input costs are correlated with either \( S \) or \( \varepsilon \), but are omitted from the equation owing to lack of data.

### 7. Energy efficiency and time costs

The model summarised in Annex A is based upon Becker’s work on the allocation of time within household production (Becker, 1965). As Binswanger (2001) has argued, time costs and the efficiency of time use have important implications for energy use in general and the rebound effect in particular. However, empirical work in this area remains in its infancy (Jalas, 2002).

For consumers, time is a necessary input to the production and enjoyment of energy services. For example, it takes time to drive from one place to another; to purchase food; to prepare a meal; to wash, dry and iron clothes and so on. The total cost of time for a particular energy service will depend upon the opportunity cost of time and the amount of time required per unit of useful work. In the household production model, the cost of time is conventionally measured by the average hourly wage for the household \( (P_W) \) and hence should vary from one household to another. The amount of time required per unit of useful work may be measured by the efficiency of time use \( (\theta) \), which depends upon the technology used. For example, a microwave oven is more time efficient than a conventional oven; a car is more time efficient than a bike;\(^\text{10}\) an aircraft is more time efficient than a ship; and so on. The relationship between useful work and time consumption for a particular energy services may then be expressed as \( S = \theta T \), while the time cost per unit of useful work may be expressed as \( P_T = P_W / \theta \). These expressions are entirely analogous to those used for energy consumption for a particular energy service (namely \( S = \varepsilon E \) and \( P_T = P_E / \varepsilon \)).

Under these assumptions, the contribution of time costs to the full cost of an energy service should be inversely proportional to the time efficiency of the relevant technology and proportional to the wage rate. Similarly, the contribution of energy costs should be inversely proportional to the energy efficiency of the relevant technology and proportional to the energy price. Consumers may be able to choose between technologies with different combinations of energy and time efficiency in the provision of a particular energy service, and also between energy services with different levels of time and energy efficiency. The relative price of time and energy should influence the direction of technological innovation and encourage higher or lower levels of time/energy efficiency for individual energy services, as well as shifts towards the development of more or less time/energy efficient services.

These considerations suggest that an increase in the cost of time (i.e. wages) relative to energy prices should induce a substitution away from time and toward energy in the production of individual services, as well as a substitution away from time-intensive services and towards energy intensive services.\(^\text{11}\) Since wages appear to have grown faster than energy prices within developed countries over the last few decades, this appears to be a fair characterisation of recent trends (Binswanger, 2001). With time costs forming a significant proportion of the total cost of many energy services, consumers and producers have sought ways to improve the time efficiency, rather than the energy efficiency, of service provision. So travel by private car has replaced walking, cycling and public transport; automatic washing machines have replaced washing by hand; fast food and ready meals have replaced traditional cooking; supermarkets (and more recently e-shopping and home delivery) have replaced the trip up the high street; email has replaced letters; and so on. Increases in aggregate energy consumption could therefore have been driven as much by the substitution of energy for time as by the overall increases in income.

The relative importance of time costs and energy costs may be expected to vary over time and between different energy services. One area where time costs are particularly important and relatively well researched is transport. For example, figures presented by Small (1992) suggest that the average time costs for US car travel were more than three times total running costs, implying that they were more than six times the total fuel costs. If the value of time is proportional to the average wage, this ratio will be higher for high-income groups and may be expected to increase over time if real incomes increase faster than energy prices. For other energy services, such as household heating, time costs may be a less significant determinant of demand. However, time costs for this service may have been much greater in the past when coal or wood fires were the norm, since time was required for preparing and lighting the fuel. In many developing countries, the time required to collect fuelwood remains an enormous burden.

\(^{10}\) Assuming no road congestion. As with energy efficiency, time efficiency is a function of the overall energy system, which could have multiple users. While congestion is given for individual decisions, it is an endogenous variable for the system as a whole.

\(^{11}\) Note that traditional consumer theory would only capture the second of these effects and that the model implies that increases in non-wage income would not encourage either type of substitution.
In those cases where technology permits a trade-off between the two, time efficiency may be represented as a function of energy efficiency (η(ε)) or vice versa (ε(η)). By taking the first of these, we may write the energy demand for a particular energy service as: \( E = s(P_0(\epsilon), P_\tau(\epsilon))/\epsilon \). This leads to an alternative definition of the rebound effect that takes into account the associated changes in time costs:

**Definition 7.**

\[
\eta_\epsilon(E) = 1 - \eta_{PS}(S) + \left[ \eta_{PS}(S)\eta_\tau(P_\tau)\eta_\epsilon(\theta) \right].
\]

Again, as compared to Definitions 3 and 4 there is an additional term in square brackets. This is the product of the elasticity of demand for useful work with respect to time costs (\( \eta_\tau(P_\tau) \)), the elasticity of time costs with respect to time efficiency (\( \eta_\epsilon(\theta) \)) and the elasticity of time efficiency with respect to energy efficiency (\( \eta(\epsilon) \)). We expect the first of these to be negative (higher time costs should reduce the demand for useful work) and the second to be positive (higher time efficiency should reduce time costs). However, the sign of the last elasticity is ambiguous: while substitution between energy and time implies that energy efficiency is negatively correlated with time efficiency, technological improvements may sometimes improve both (e.g. microwave ovens). However, in many cases greater energy (time) efficiency is likely to be achieved at the expense of lower time (energy) efficiency (i.e. \( \theta \) and \( \epsilon \) will be negatively correlated). For example, a sports car is less energy efficient than a Smart car; aircraft are less energy efficient than ships; washing machines are less energy efficient than handwashing; and so on. In these circumstances, the resulting increase in time costs will offset the saving in energy costs leading to a smaller rebound effect. For example, while greater fuel efficiency may make driving cheaper, consumers may not be willing to spend the time driving greater distances. This again suggests that empirical estimates based upon Definitions 3 and 4 and relying primarily upon historical or cross-sectional variations in energy prices may overestimate the magnitude of the rebound effect.

As with capital costs, the size of this upward bias will depend upon the relative magnitude of the different elasticities. One notable implication is that rebound effects from improved energy efficiency may be expected to decrease over time, since GDP growth should increase average wages and make time costs relatively more important in the total cost of energy services. One of the few studies to show evidence for this is Small and Van Dender (Small and Van Dender, 2005), although their methodology was subsequently criticised by Harrison et al. (2005).

If improvements in energy efficiency are associated with changes in both time and capital costs, the appropriate expression for the rebound effect becomes:

**Definition 8.**

\[
\eta_\epsilon(E) = 1 - \eta_{PS}(S) - \left[ \eta_{PS}(S)\eta_\tau(P_\tau)\eta_\epsilon(\theta) \right] + \left[ \eta_{PS}(S)\eta_\tau(P_\tau)\eta_\epsilon(\theta) \right].
\]

As pointed out by Binswanger (2001), the analogy between time and energy efficiency also suggests that there should be a parallel rebound effect with respect to time. Since improvements in the time efficiency associated with a particular service lower the cost of that service, there should be a corresponding increase in service demand that will offset the potential time savings. Again, transport provides a particularly good example: the potential time savings from faster modes of transport may be partly or wholly taken back by traveling greater distances. Similar patterns are likely to apply to other services (e.g. washing clothes more often), but may be less noticeable if time costs form a smaller proportion of total costs, or if the assumptions of the simple Becker model (e.g. no joint production) do not apply.

The rebound effect with respect to time may be defined as an efficiency elasticity (\( \eta(T) = \eta(S) - 1 \)) or as a price elasticity (\( \eta(T) = \eta(S) - 1 \)) in a similar manner to the conventional rebound effect. Empirical investigation of this effect would similarly need to take into account the potential correlation between improvements in time efficiency and other input costs (including capital and energy costs); and the potential endogeneity of time efficiency (e.g. consumers may choose a more time efficient technology if they anticipate a high demand for the service). But in the absence of good data on time use patterns and time efficiency, such considerations remain academic.

Both the substitution of energy for time in the production and consumption of energy services, and the subsequent rebound effect with respect to time should act to increase overall energy consumption. Indeed, it is possible that these processes have had a more important influence upon aggregate energy consumption than the conventional rebound effect with respect to energy efficiency. Moreover, if wages continue to increase faster than energy prices, the substitution of energy for time may be expected to increase in importance, while the conventional rebound effect decreases in importance. To date, however, analytical and political attention has focused disproportionately on the latter.

### 8. Summary

This paper has sought to clarify and bring together a number of definitions of the direct rebound effect and identify their underlying assumptions. It has clarified the relationship between the ‘engineering’ definition of the direct rebound effect as an efficiency elasticity and the more common definition in the economic literature as a price elasticity. It has discussed a number of factors that need to be taken into account when developing such empirical estimates and emphasised the trade-offs between both the different categories of input costs and between useful work and other attributes of an energy service. It has also shown how different measures of useful work (together with differing ways of decomposing those measures) may lead to different conclusions regarding the nature and size of direct rebound effects.

Many empirical estimates of the direct rebound effect are based upon price elasticities and rely primarily upon historical or cross-sectional variations in energy prices. The paper has argued that such studies could potentially overestimate the magnitude of the effect. Factors contributing to this include: the asymmetry of price elasticity estimates; the anticipated positive correlation between energy efficiency and other categories of input costs, notably capital costs; the role of price induced efficiency improvements; the endogeneity of energy efficiency; and the anticipated negative correlation between energy...
efficiency and time efficiency. Different studies address these factors in different ways and to a greater or lesser extent, with some of the best examples being recent US studies of travel by private car (Greene et al., 1999b; Small and Van Dender, 2005). Those studies that use the own-price elasticity of energy demand as a proxy for the direct rebound effect appear to be particularly flawed.

The existing evidence base for direct rebound effects is methodologically diverse and limited in scope, with most studies being confined to car travel and household heating in the US. While data limitations provide constraints, there should be scope for research using alternative definitions of useful work (e.g. tonne kilometers for passenger travel), encompassing a greater range of energy services (e.g. washing machines and refrigerators), incorporating both short and long term effects (e.g. distinguishing between changes in the number, capacity and utilisation of conversion devices) and including energy services in developing countries (where rebound effects may be larger). Perhaps the greatest area of neglect is the time costs associated with energy service provision, due largely to the lack of adequate data in this area. However, the substitution of energy for time in the provision of energy services, together with the parallel ‘rebound effect with respect to time’ are likely to be important drivers of increases in aggregate energy consumption. Both of these deserve further research.

Acknowledgements

This research was funded by the UK Research Councils as part of the Technology and Policy Assessment (TPA) function of the UK Energy Research Centre (UKERC). An earlier version of this paper was presented at the 29th IAEE International Conference ‘Securing energy in insecure times’, held in Potsdam, Germany on the 7–10th June 2006. The authors gratefully acknowledge comments from Harry Saunders, John Feather, Blake Alcott, Brenda Boardman, Mark Barrett and Ed Steinmuller as well as from participants at the IAEE meeting and two anonymous referees. A debt is also owed to Lorna Greening and David Greene for their previous synthesis of empirical work in this area (Greening and Greene, 1998). The usual disclaimers apply.

Annex A — The household production model

In Becker’s household production model, individual households are assumed to produce energy services (ES) by combining energy (E), capital (K) and other market goods (O), together with some of the household’s own time (T). To reflect this, the production function for energy service i may be written as:

\[
ES_i = ES_i[E_i, K_i, O_i, T_i; A_i].
\]  

(A.1)

If a household’s utility is assumed to depend solely upon these services, the utility function becomes:

\[
U = u[ES_1, ES_2, ES_3, ..., ES_n].
\]

(A.2)

The household may be assumed to be subject to the following income constraint:

\[
V + TWP_W \geq \sum_{i=1}^{n} (P_E E_i + P_O O_i + \delta_K K_i).
\]

(A.3)

Where \( V \) represents non-wage income; \( P_W \) represents the average wage rate; \( T_W \) represents the time spent in the labour market; \( P_E \) and \( P_O \) represent the unit price of energy and other goods respectively; and \( \delta_K \) represents a discount factor (so \( \delta_K = \delta_K(A) \)) gives the annualised capital costs). Households will also be subject to a second constraint on their available time:

\[
T = T_W + \sum_{i=1}^{n} T_i.
\]

(A.4)

Where \( T_i \) represents the time spent in producing services \( S \). Becker (1965) argued that, since money and time are partly interchangeable through decisions on \( T_W \), the income and time constraints can be collapsed into a single constraint. By substituting \( T_W = T - \sum T_i \) into the budget constraint and rearranging, we obtain:

\[
V + P_W T \geq \sum_{i=1}^{n} (P_E E_i + P_O O_i + \delta_K K_i + P_W T_i).
\]

(A.5)

Versions of Becker’s ‘household production’ model form the basis of a substantial volume of empirical research (Juster and Stafford, 1991; Gronau, 1997). This includes numerous applications to energy use, although these studies frequently (and importantly) neglect the time inputs to energy services (Dinan, 1987; Willet and Naghshpour, 1987; Davis, 2004). The model rests upon a set of behavioural and other assumptions that may be criticised on a variety of grounds (Pollack and Wachter, 1975; Juster and Stafford, 1991). Nevertheless, it offers a number of advantages over conventional models of household demand (especially for energy) and predictions from the model appear broadly confirmed by empirical research (Juster and Stafford, 1991).

Annex B — Mathematical derivations

Derivation of Definition 1

Given \( S = rE \)

\[
\eta_r(E) = \frac{\partial (S/E)}{\partial r} \left( \frac{E}{S/r} \right) = \left( -S \frac{1}{r^2} + \frac{1}{r} \frac{\partial S}{\partial r} \right) \left( \frac{E^2}{S} \right) = \frac{\partial S}{\partial r} - 1.
\]

Or: \( \eta_r(E) = -\eta_r(S) - 1. \)

\(12\) Including: the assumption that each market good or allocation of time is dedicated to the production of a single service; the notion that households are indifferent to the allocation of time, except as an input into the production of services; difficulties in defining what a service actually is (e.g. travel by car for a visit or the visit itself); the neglect of the fact that utility may be a function of producing as well as consuming a service; the implicit assumption of constant returns to scale in production; the difficulty in operationalising the model; the lack of good data on time use patterns; and the usual difficulties associated with models that assume ‘hyper-rational’, utility maximising individuals.
Derivation of Definition 2

Given \( S = \varepsilon E \) and \( S = NO + CAP + UTIL \)

\[
\eta_i(E) = \frac{\varepsilon}{E} \left[ \frac{(NO + CAP + UTIL)}{\varepsilon} \right] + 1 \frac{(NO + CAP + UTIL)}{\varepsilon} \frac{\partial Util}{\partial E} \\
+ \left( \frac{(NO + UTIL)}{\varepsilon} \right) \frac{\partial Util}{\partial E} \frac{\partial Util}{\partial S} \frac{\partial Util}{\partial S} \frac{\partial Util}{\partial S} \right] \]

Substituting \( E = (NO + UTIL)/\varepsilon \) and cancelling terms:

\[
\eta_i(E) = -1 + \left( \frac{\varepsilon}{E} \right) \frac{\partial Util}{\partial E} \frac{\partial Util}{\partial S} \frac{\partial Util}{\partial S} \frac{\partial Util}{\partial S} \right] \]

Or: \( \eta_i(E) = (\eta_i(CAP) + \eta_i(Util)) - 1 \).

Derivation of Definition 3

Given \( E = S(P_S) / \varepsilon \) and \( S = P_S / \varepsilon \) and assuming that \( P_S \) is exogenous, we have:

\[
\eta_i(S) = \frac{\partial S}{\partial P_S} \frac{S}{P_S} \frac{\varepsilon}{E} \frac{\partial S}{\partial P_S} \frac{\partial S}{\partial P_S} \frac{\partial S}{\partial P_S} \right] \]

But if energy efficiency is held constant the above relationship becomes:

\[
\eta_i(S) = \frac{\partial E}{\partial P_S} \frac{P_S}{E} = \eta_i(E) \]

Or: \( \eta_i(E) = -\eta_i(S) - 1 \).

Derivation of Definition 4

Including the capital costs of new equipment \( P_D \), the basic identity becomes:

\[
E = S|P_S|/\varepsilon \]

Taking derivatives with respect to energy efficiency, we have:

\[
\frac{\partial E}{\partial \varepsilon} = \frac{\partial S}{\partial \varepsilon} \frac{P_S}{E} + \frac{1}{\varepsilon} \frac{\partial S}{\partial P_S} \frac{\partial P_S}{\partial \varepsilon} \frac{\partial E}{\partial S} \frac{\partial \varepsilon}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial \varepsilon} \right] \]

Multiplying through by \( \varepsilon / E \) to obtain \( \eta_i(E) \):

\[
\frac{\partial E}{\partial \varepsilon} = \frac{S}{\varepsilon} \frac{P_S}{E} \frac{\partial S}{\partial \varepsilon} \frac{\partial P_S}{\partial \varepsilon} \frac{\partial E}{\partial S} \frac{\partial \varepsilon}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial \varepsilon} \right] \]

Multiplying numerator and denominator of the last term with \( P_S \), we have:

\[
\frac{\partial E}{\partial \varepsilon} = \frac{S}{\varepsilon} \frac{P_S}{E} \frac{\partial S}{\partial \varepsilon} \frac{\partial P_S}{\partial \varepsilon} \frac{\partial E}{\partial S} \frac{\partial \varepsilon}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial \varepsilon} \right] \]

Or: \( \eta_i(E) = \frac{-1 - \eta_i(S) + \eta_i(P_S)}{\eta_i(P_S)} \).

Derivation of Definition 6

If energy efficiency depends upon energy prices, the basic identity can be written as follows:

\[
E = S \frac{\partial S}{\partial P_S} \frac{\partial P_S}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial \varepsilon} \right] \]

Use the product and chain rules to differentiate this with respect to energy commodity prices:

\[
\frac{\partial E}{\partial P_S} \frac{\partial \varepsilon}{\partial \varepsilon} = \frac{S}{\varepsilon} \frac{P_S}{E} \frac{\partial S}{\partial P_S} \frac{\partial P_S}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial \varepsilon} \right] \]

Multiplying both sides by \( P_S / E \) to switch into elasticity forms:

\[
\frac{\partial E}{\partial P_S} \frac{\partial \varepsilon}{\partial \varepsilon} = \frac{S}{\varepsilon} \frac{P_S}{E} \frac{\partial S}{\partial P_S} \frac{\partial P_S}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial \varepsilon} \right] \]

Expressing each term as an elasticity, we obtain:

\[
\eta_i(E) = \eta_i(S) - \eta_i(e)(1 + \eta_i(S)) \]

Or alternatively: \( \eta_i(S) = \frac{-1 - \eta_i(e)}{\eta_i(S)} \).

Derivation of the relative magnitude of price elasticities

Starting with the identity \( E = S|P_S|/\varepsilon \), the energy cost elasticity of the demand for useful work may be expressed as:

\[
\eta_i(S) = \frac{S}{\varepsilon} \frac{P_S}{E} \frac{\partial S}{\partial P_S} \frac{\partial P_S}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial \varepsilon} \right] \]

Or:

\[
\eta_i(S) = \eta_i(S) - \eta_i(e) \]

We expect that \( \eta_i(\varepsilon) \geq 0 \) (higher costs for useful work encourages higher energy efficiency). In contrast, we expect that \( \eta_i(e) \leq 0 \) (higher prices reduce demand). Hence we expect that:

\[
|\eta_i(E)| \geq |\eta_i(S)| \]

By a very similar process we can show:

\[
|\eta_i(E)| \geq |\eta_i(S)| \]

And hence we can argue that:

\[
|\eta_i(E)| \geq |\eta_i(S)| \]
Rearranging Definition 6 we have:

$$\eta_\varepsilon(E) = \eta_\varepsilon(S) [1 - \eta_\varepsilon(\varepsilon)] - \eta_\varepsilon(\varepsilon).$$

In most cases we would expect $1 \geq \eta_\varepsilon(\varepsilon) \geq 0$ and $0 \geq \eta_\varepsilon(S) \geq -1$. This implies that:

$$\left| \eta_\varepsilon(S) \right| \leq \left| \eta_\varepsilon(E) \right|.$$

Combining the above three relationships, we obtain:

$$\left| \eta_\varepsilon(S) \right| \leq \left| \eta_\varepsilon(S, E) \right| \leq \left| \eta_\varepsilon(E) \right|.$$

**Derivation of Definition 7**

Including time costs and assuming time efficiency ($\theta$) we have:

$$E = s[P_3(\varepsilon), P_1(\theta(\varepsilon))]/\varepsilon.$$  

Taking derivatives with respect to energy efficiency, we have:

$$\frac{\partial E}{\partial \varepsilon} = \frac{S}{\varepsilon} + \frac{1}{\varepsilon} \left[ \frac{\partial S}{\partial P_3} \frac{\partial P_3}{\partial \varepsilon} + \frac{\partial S}{\partial P_1} \frac{\partial P_1}{\partial \varepsilon} \right].$$

$$= \frac{S}{\varepsilon} - \frac{1}{\varepsilon^2} - \frac{1}{\varepsilon} \frac{\partial S}{\partial P_3} \frac{\partial P_3}{\partial \varepsilon}.$$

Multiply through by $\varepsilon/E$ to obtain $\eta_\varepsilon(E)$:

$$\frac{\partial E}{\partial \varepsilon} = \frac{S}{\varepsilon} - \frac{1}{\varepsilon^2} - \frac{1}{\varepsilon} \frac{\partial S}{\partial P_3} \frac{\partial P_3}{\partial \varepsilon}. $$

Multiply the third term by ($\theta P_3/\theta P_1$) and rearrange:

$$\frac{\partial E}{\partial \varepsilon} = -\frac{S}{\varepsilon} \frac{\partial S}{\partial P_3} \frac{\partial P_3}{\partial \varepsilon} + \left( \frac{\partial P_3}{\partial \theta} \frac{\partial \theta}{\partial P_1} \right) \frac{\varepsilon}{\partial \varepsilon}.$$  

Or: $\eta_\varepsilon(E) = -1 - \eta_\varepsilon(S) + [\eta_\varepsilon(S) \eta_\varepsilon(P_3) \eta_\varepsilon(\theta)]$.

**REFERENCES**


Empirical estimates of the direct rebound effect: A review

Steve Sorrell a, *, John Dimitropoulos b, Matt Sommerville c

a Sussex Energy Group, SPRU (Science & Technology Policy Research), Freeman Centre, University of Sussex, Falmer, Brighton BN1 9QE, UK
b Mott MacDonald, Brighton, UK
c Sussex Energy Group, SPRU (Science & Technology Policy Research), Freeman Centre, University of Sussex, Falmer, Brighton BN1 9QE, UK

A R T I C L E   I N F O

Article history:
Received 15 June 2008
Accepted 17 November 2008
Available online 24 January 2009

Keywords:
Rebound effect
Take-back
Energy efficiency

A B S T R A C T

Improvements in energy efficiency make energy services cheaper, and therefore encourage increased consumption of those services. This so-called direct rebound effect offsets the energy savings that may otherwise be achieved. For example, consumers may choose to drive further and/or more often following the purchase of a fuel-efficient car because the operating cost per kilometre has fallen. Similarly, consumers may choose to heat their homes for longer periods and/or to a higher temperature following the installation of loft insulation, because the operating cost per square metre has fallen. The extent to which this occurs may be expected to vary widely from one energy service to another, from one circumstance to another and from one time period to another. But any increase in energy service consumption will reduce the ‘energy savings’ achieved by the energy efficiency improvement. In some circumstances, it could offset those savings altogether—an outcome that has been termed ‘backfire’.

Direct rebound effects are the most familiar and widely studied component of the overall or economy-wide rebound effect (Sorrell, 2007) which also involves various indirect effects (for example, the energy associated with other goods and services whose consumption has increased as a result of the energy efficiency improvement). Beginning with Khazzoom (1980), there have been a series of estimates of the direct rebound effect for different energy services (Greening and Greene, 1998). These studies are extremely diverse in terms of the definitions, methodological approaches and data sources used. Also, despite growing research activity, the evidence remains sparse, inconsistent and largely confined to a limited number of consumer energy services in the United States—notably personal automobile transport and household heating. The main reason for this is the lack of suitable data sources for other types of energy service in other sectors and countries. In addition, interpretation of the evidence is greatly hampered by the use of competing definitions, measures, terminology and notation. Many studies do not mention the direct rebound effect at all, but nevertheless provide elasticity estimates that may, under certain assumptions, be used as proxy measures of that effect. Taken together, these features inhibit understanding of the direct rebound effect and the appropriate methodological approach to estimating its magnitude in different circumstances, as well as making it difficult to identify the relevance of particular studies.

This paper provides an overview of the methodological approaches to estimating direct rebound effects and reviews the evidence that is currently available. It updates an earlier review by Greening et al. (2000) and seeks to clarify a number of issues that were raised therein. The underlying research is reported in detail in Sommerville and Sorrell (2007) and Sorrell and Dimitropoulos (2007a). The paper focuses entirely on energy services in the household sector, since this is where practically all of the research has been undertaken. As a result, the conclusions do not provide guidance on the magnitude of direct rebound effects in other sectors, nor on the economy-wide rebound effect, which is fully discussed by Sorrell (2007) and Sorrell and Dimitropoulos (2007c).

Section 1 describes the operation of the direct rebound effect, highlighting some key issues concerning the measurement of this effect and the conditions under which it may be expected to be larger or smaller. Sections 2 and 3 describe the...
2. Understanding the direct rebound effect

Energy services such as heating and lighting are provided through energy systems that involve particular combinations of capital equipment, labour, materials and marketable energy commodities such as electricity. The relevant systems may include primary conversion equipment such as boilers, secondary conversion equipment such as radiators, equipment for distributing energy and manual or electronic controls. For space heating and lighting the relevant energy systems may also include building fabric, thermal insulation, ventilation systems and glazing. The energy efficiency of such systems \((e)\) is defined as the ratio of useful energy outputs \((S)\) to energy inputs \((E)\) and may be influenced by a variety of factors other than the thermodynamic efficiency of particular conversion equipment. Different measures of energy efficiency can be developed for different system boundaries \((e.g.\) the boiler or the house\) and the magnitude of such indicators will depend upon how the energy inputs and outputs are defined and measured.

Useful energy outputs \((S)\) may be measured by a variety of thermodynamic or physical indicators, the appropriate choice of which will depend upon the system under consideration, the purpose of the analysis and the availability of the relevant data \((Patterson, 1996)\). For example, the energy service delivered by passenger cars may be measured in terms of vehicle kilometres, passenger kilometres or \((\text{rather unconventionally})\) tonne kilometres.

Energy services may also have broader attributes that may be combined with useful energy outputs in a variety of ways. For example, all cars deliver passenger kilometres, but they may vary widely in terms of speed, comfort, acceleration and prestige. Consumers and producers may therefore make trade-offs between useful energy outputs and other attributes of an energy service: between energy, capital and other market goods in the production of an energy service\(^1\); and between different types of energy service.

The cost of energy services \((P_S)\) may be defined as: \(P_S = P_E/e\), where \(P_E\) is the price of energy inputs. This, however, is only one component of the overall cost of providing an energy service \((P_C)\), where annualised capital, maintenance and time costs must also be included. Energy efficiency improvements reduce the energy-related and, hence, the overall cost of a service. Over time, this may encourage an increase in the number of energy conversion devices, their average size, their average utilisation and/or their average load factor. For example, people may buy more cars, buy larger cars, drive them further and/or share them less. Similarly, people may buy more washing machines, buy larger machines, use them more frequently and/or reduce the average load. The relative importance of these variables may be expected to vary widely between different

---

\(^1\) See Sorrell and Dimitropoulos \((2007a)\) for an application of Becker's ‘household production model’ to energy services.
direct rebound effects may be smaller if energy efficient equipment is more expensive than less-efficient alternatives, because the discounted lifetime savings will be less (Henly et al., 1988). Since consumers typically have high implicit discount rates, the availability of such equipment may not encourage an increase in the number of units purchased, or their average size. However, once purchased, such equipment may be expected to have a higher utilisation, owing to their lower short-run marginal costs. In practice, many types of equipment appear to have both improved in energy efficiency over time and fallen in overall cost relative to income.

Even if energy efficiency improvements are not associated with changes in capital or other costs, certain types of direct rebound effect may be constrained by the real or opportunity costs associated with increasing demand. Two examples are the opportunity cost of space (e.g. increasing refrigerator size may not be the best use of available space) and the opportunity cost of time (e.g. driving longer distances may not be the best use of available time). However, space constraints may become less important over time if technological improvements reduce the average size of conversion devices per unit of output or if rising incomes lead to an increase in average living space (e.g. compare refrigerator sizes in the US and the UK) (Wilson and Boehland, 2005). In contrast, the opportunity cost of time should increase with rising incomes. The direct rebound effect for a particular energy service may therefore vary between households and over time and may be influenced by a large number of variables.

3. The quasi-experimental approach

One approach to estimating direct rebound effects relies upon measuring the demand for the energy service before and after an energy efficiency improvement: for example, measuring the change in heat output following the installation of a fuel-efficient boiler. The demand for the energy service before the energy efficiency improvement could be taken as an estimate for what demand ‘would have been’ in the absence of the improvement. However, various other factors may also have changed the demand for the energy service which need to be controlled for (Frodel and Schmidt, 2005; Meyer, 1995).

Since it can be very difficult to measure energy service demand, an alternative approach is to measure the change in energy inputs (e.g. the fuel consumed by the boiler). But to estimate direct rebound effects, this needs to be compared with a counterfactual estimate of energy consumption that has at least two sources of error, namely: (a) the energy consumption that would have occurred without the energy efficiency improvement; and (b) the energy consumption that would have occurred following the energy efficiency improvement had there been no behavioural change. The first of these gives an estimate of the energy savings from the energy efficiency improvement, while the second isolates the direct rebound effect. Estimates for the latter can be derived from engineering models, but these frequently require data on the circumstances of individual installations and are prone to error.

Both of these approaches are relatively rare, owing in part to measurement difficulties. There are few peer-reviewed studies and nearly all of these focus on household heating (Sommerville and Sorrell, 2007). The methodological quality of most of these studies is relatively poor, with the majority using simple before–after comparisons, without the use of a control group or explicitly controlling for confounding variables. This is the weakest methodological strategy and prone to bias (Frodel and Schmidt, 2005; Meyer, 1995). Also, many studies are vulnerable to selection bias, since households choose to participate rather being randomly assigned (Hartman, 1988). Other weaknesses include small sample sizes, a failure to present the error associated with estimates, large variation in the relevant independent variable both within and between studies (e.g. households receiving different types of energy efficiency measure, or combination of measures) and monitoring periods that are too short to capture long-term effects. In the case of household heating, there is also persistent confusion between

- shortfall, the difference between actual savings in energy consumption and those expected on the basis of engineering estimates;2
- temperature take-back, the change in mean internal temperatures following the energy efficiency improvement, or the reduction in energy savings associated with that change; and
- behavioural change, the proportion of the change in internal temperature that derives from adjustments of heating controls and other variables by the user (e.g. opening windows), or the reduction in energy savings associated with those changes.

Unfortunately, different studies using different terms for the above concepts as well as the same term for different concepts. Typically, only a portion of temperature take-back is due to behavioural change, with the remainder being due to physical and other factors (Sanders and Phillipson, 2006). Similarly, only a portion of shortfall is due to temperature take-back, with the remainder being due to poor engineering estimates of potential savings, inadequate performance of equipment, deficiencies in installation and so on. Hence, behavioural change is one, but not the only (or necessarily the most important) explanation of temperature take-back and the latter is one, but not the only explanation of shortfall. Direct rebound effects are normally interpreted as behavioural change, but it may be misleading to interpret this solely as a rational response to lower heating costs, partly because energy-efficiency improvements may change other variables (e.g. airflow) that also encourage behavioural responses. Also, measures of temperature change may be difficult to translate into estimates of reduced energy savings because of the non-linear and context-specific relationship between energy consumption and internal temperature. Isolating direct rebound effects from such studies can, therefore, be challenging.

4. The econometric approach

A more common approach to estimating direct rebound effects is through the econometric analysis of secondary data sources that include information on the demand for energy, the relevant energy service and/or the energy efficiency of that service. This data can take a number of forms (e.g. cross-sectional, time-series, panel) and apply to different levels of aggregation (e.g. household, region, country). Such studies typically estimate elasticities, meaning the percentage change in one variable following a percentage change in another, holding the other measured variables constant. If time-series data is available, an estimate can be made of short-run elasticities, where the stock of conversion devices is assumed to be fixed, as well as long-run elasticities where it is variable. Cross-sectional data are usually assumed to provide estimates of long-run elasticities.3

---

2 The UK government uses the misleading term ‘comfort factor’ to describe the shortfall in energy savings following improvements in their thermal efficiency of housing. However, a recent review of UK studies (Sanders and Phillipson, 2006) uses the term comfort factor to describe temperature take-back and reduction factor to describe shortfall.

3 This assumes that demand is in equilibrium at the point of observation, or at least that the relationship between the explanatory variables has been approximately constant in the recent past.
Depending upon data availability, the direct rebound effect may be estimated from one of two energy-efficiency elasticities:\footnote{The rationale for the use of these elasticities, and the relationship between them, is explained in detail in Sorrell and Dimitropoulos, 2007a.}

- $\eta_{j}(E)$: the elasticity of demand for energy ($E$) with respect to energy efficiency ($e$)
- $\eta_{j}(S)$: the elasticity of demand for energy services ($S$) with respect to energy efficiency (where $S = E^{j}$)

$\eta_{j}(S)$ is generally taken as a direct measure of the rebound effect. From the relationship $S = E^{j}$ it can easily be shown that: $\eta_{j}(E) = \eta_{j}(S) - 1$ (Sorrell and Dimitropoulos, 2007a). Hence, the actual saving in energy consumption will only equal the predicted saving from engineering calculations when the demand for energy services remains unchanged following an energy-efficiency improvement (i.e. when $\eta_{j}(S) = 0$). In these circumstances, an $\%$ improvement in energy efficiency should lead to an $\%$ reduction in energy consumption (i.e. $\eta_{j}(E) = -1$). A positive rebound effect implies that $\eta_{j}(S) > 0$ and $0 > \eta_{j}(E) > -1$, while backfire implies that $\eta_{j}(S) > 1$ and $\eta_{j}(E) = 0$. The rebound effect is normally quoted in percentage terms, so a rebound effect of 20\% means that $\eta_{j}(S) = 0.2$, $\eta_{j}(E) = -0.8$ and that 20\% of the potential energy savings are ‘taken back’ as a result of the increased demand for energy services.

But instead of using $\eta_{j}(E)$ or $\eta_{j}(S)$, most studies estimate the rebound effect from one of three price elasticities, namely:

- $\eta_{j}(P_{S})$: the elasticity of demand for energy services with respect to the energy cost of energy services ($P_{S}$)
- $\eta_{j}(S)$: the elasticity of demand for energy services with respect to the price of energy ($P_{E}$)
- $\eta_{j}(E)$: the elasticity of demand for energy with respect to the price of energy.

Where $P_{S} = P_{E}/E$. Under certain assumptions, the negative of either $\eta_{j}(P_{S})$, $\eta_{j}(S)$ or $\eta_{j}(E)$ can be taken as an approximation to $\eta_{j}(S)$, and hence may be used as a measure of the direct rebound effect (Sorrell and Dimitropoulos, 2007a). The use of price elasticities in this way implicitly equates the direct rebound effect to a behavioural response to the lower cost of energy services. It therefore ignores any other reasons why the demand for energy services may change following an improvement in energy efficiency (e.g. households responding to changes in ventilation patterns following insulation improvements).

The choice of the elasticity measure will depend in part upon data availability.\footnote{In the case of personal automotive transport, for example, the elasticities could correspond to: $\eta_{j}(E)$—the elasticity of the demand for motor-fuel (for passenger cars) with respect to kilometres per litre; $\eta_{j}(S)$ the elasticity of the demand for vehicle kilometres with respect to kilometres per litre; $\eta_{j}(E)$ the elasticity of the demand for vehicle kilometres with respect to the cost per kilometre; $\eta_{j}(S)$ the elasticity of the demand for motor-fuel with respect to the price of motor-fuel; and $\eta_{j}(E)$ the elasticity of the demand for motor-fuel with respect to the price of motor-fuel.} Generally, data on energy consumption ($E$) and energy prices ($P_{E}$) is both more available and more accurate than data on energy services ($S$) and energy efficiency ($e$). Also, even if data on energy efficiency is available, the amount of variation is typically limited, with the result that estimates of either $\eta_{j}(E)$ or $\eta_{j}(S)$ can have a large variance. In contrast, estimates of $\eta_{j}(P_{S})$ may have less variance owing to significantly greater variation in the explanatory variable ($P_{S}$). This is because the energy cost of energy services depends upon the ratio of energy prices to energy efficiency ($P_{S} = P_{E}/E$) and most data sets include considerable cross-sectional or longitudinal variation in energy prices.

In principle, rational consumers should respond in the same way to a decrease in energy prices as they do to an improvement in energy efficiency (and vice versa), since these should have an identical effect on the energy cost of energy services ($P_{S}$). However, changes in energy efficiency may be correlated with changes in other input costs, while changes in energy prices may not. If these are not controlled for, estimates of the direct rebound effect that are based upon $\eta_{j}(S)$ and which rely upon variations in energy prices could be biased (see Section 8).

Estimates of $\eta_{j}(P_{S})$ are largely confined to personal automotive transportation, household heating and space cooling, where proxy measures of energy services are most readily available. These services form a significant component of household energy consumption in OECD countries and demand for them may be expected to be relatively price responsive. There are very few estimates of $\eta_{j}(S)$ for other consumer energy services and practically none for producers. Furthermore, the great majority of studies refer to the United States.

In many cases, data on energy efficiency is either unavailable or inaccurate. In these circumstances, the direct rebound effect may be estimated from $\eta_{j}(P_{S})$ or $\eta_{j}(E)$. But this is only valid if: first, consumers respond in the same way to a decrease in energy prices as they do to an improvement in energy efficiency (and vice versa); and second, energy efficiency is unaffected by changes in energy prices (i.e. $\eta_{j}(E) = 0$). Both these assumptions are likely to be flawed, but the extent to which this leads to biased estimates of the direct rebound effect may vary widely from one energy service to another and between the short and long term.

Under certain assumptions, the own-price elasticity of energy demand ($\eta_{j}(P_{S})$) for a particular energy service can be shown to provide an upper bound for the direct rebound effect (Sorrell and Dimitropoulos, 2007a).\footnote{Sorrell and Dimitropoulos (2007b) show that the following relationship may be expected to hold: $|\eta_{j}(S)| < |\eta_{j}(P_{S})| < |\eta_{j}(E)| < \eta_{j}(E)$. See also} As such, the voluminous literature on energy price elasticities may be used to place some bounds on the likely magnitude of the direct rebound effect for different energy services in different sectors. This was the approach taken by Khazzoom (1980), who pointed to evidence that the long-run own-price elasticity of energy demand for water heating, space heating and cooking exceeded (minus) unity in some circumstances, implying that energy-efficiency improvements for these services could lead to backfire (Taylor et al., 1977). However, reviews of this literature generally suggest that energy demand is inelastic in the majority of sectors in OECD countries (i.e. $\eta_{j}(P_{S}) < 1$) (Dahl, 1993, 1994; Dahl and Sterner, 1991; Espey and Espey, 2004; Espey, 1998; Graham and Glaister, 2002; Hanley et al., 2002). The implication is that the direct rebound effect is unlikely to lead to backfire within OECD countries.

For the purpose of estimating rebound effects, estimates of $\eta_{j}(E)$ are most useful when the energy demand in question relates to a single energy service, such as refrigeration. They are less useful when (as is more common) the measured demand derives from a collection of energy services, such as household fuel or electricity consumption. In this case, a large value for $\eta_{j}(E)$ may suggest that improvements in the ‘overall’ efficiency of fuel or electricity use will lead to large direct rebound effects (and vice versa), or that the direct rebound effect for the energy services that dominate fuel or electricity consumption may be large. However, a small value for $\eta_{j}(E)$ would not rule out the possibility of large direct rebound effects for individual energy services.

Whatever their scope and origin, estimates of price elasticities should be treated with caution. Aside from the difficulties of estimation, behavioural responses are contingent upon technical, institutional, policy and demographic factors that vary widely...
5. Estimates for personal transport

By far the best studied area for the direct rebound effect is personal automotive transport. Most studies refer to the US, which is important since fuel prices, fuel efficiencies and residential densities are lower than in Europe, car ownership levels are higher and there is less scope for switching to alternative transport modes.

In principle, estimates of the own-price elasticity of gasoline consumption for personal transport \( \eta_p(E) \) should provide an upper bound for the direct rebound effect for this energy service. The numerous empirical estimates of \( \eta_p(E) \) have been comprehensively reviewed by Goodwin (1992), Espey (1996, 1998), Hanley et al. (2002) and Graham and Glaison (2004). Using the mean estimates from these meta-analyses suggests an upper bound of for the short-term direct rebound effect of 20–25%, increasing to 80% over the long-term (Sorrell and Dimitropoulos, 2007b). However, there is a large variance in the results of different studies and the upper bound is likely to be significantly larger than the actual effect since the use of \( \eta_p(E) \) neglects the effect of fuel prices on vehicle fuel efficiency.

More accurate estimates of the direct rebound effect for personal transport may be obtained from studies estimating \( \eta_p(E), \eta_p(S) \) or \( \eta_p(S) \), but these vary considerably in terms of the data used and specifications employed. Sorrell and Dimitropoulos (2007a) review 17 such studies, the main features and results of which are summarised in Tables A2–A4. While all of these studies use distance travelled as a measure of the energy service \( S \), this may either be measured in absolute terms or normalised to the number of adults, licensed drivers, households or vehicles (Sorrell and Dimitropoulos, 2007a). The relevant estimates may be expected to differ as a result.

The seven studies using aggregate time-series and cross-sectional data (Table A2) estimate the long-run direct rebound effect for personal automotive transport to be somewhere between 5% and 30%. While there is disagreement over the appropriate specification, particularly in relation to the appropriate treatment of serial correlation and lagged dependent variables, the limited number of data points available makes it difficult to settle the issue from this type of data alone (Greene, 1992; Jones, 1993). Also, since these studies refer solely to the US and use relatively old data, they cannot be used to investigate whether direct rebound effects have declined over time or whether they are substantially different in other countries.

The four studies using cross-country and regional panel data (Table A3) provide substantially more observations than time-series or cross-sectional data and thereby provide a more robust basis for estimating the direct rebound effect. Three of these studies are very carefully done and incorporate some important methodological innovations. The cross-country study of Johansson and Schipper (1997) gives results at the high end of the range in the literature (i.e. a best guess for the long-run direct rebound effect of 30%), while both Haughton and Sarkar (1996) and Small and Van Dender (2005, 2007) converge on a long-run value of 22% for the US. Small and van Dender’s study provides a strong indication that the direct rebound effect declines over time as incomes increase, which is consistent with theory. However, their model has a number of weaknesses (Box 1) and the suggestion of declining rebound effects is not supported by the meta-analysis by Hanley et al. (2002) of other studies in this area.

The five studies using disaggregate data sources (Table A4) provide less-consistent estimates of the direct rebound effect and several of these estimates are surprisingly high. While disaggregate data avoids some of the measurement difficulties reported by Schipper et al. (1993), the greater complexity of the models can create some difficulties in interpretation. It is notable that three of the studies use data from the US Consumer Expenditure Survey taken from overlapping periods, but nevertheless produce estimates of the direct rebound effect that range from 0% to 87%. This diversity suggests that the results from disaggregate studies should be interpreted with caution.

Of the disaggregate studies, Greene et al. (1999) provide the most careful investigation of the direct rebound effect, although their model does not allow for long-term changes in the vehicle stock. This study also produces an estimate of the long-run direct rebound effect (23%) which is consistent with the results of aggregate studies. While the study from Frondel et al. (2008) is suggestive of large rebound effects in Germany, the lack of comparable European studies and the inconsistency between these results and those of aggregate estimates of fuel price elasticities provide an insufficient basis to conclude that direct rebound effects are larger in Europe.

On the basis of this review, we can conclude that the long-run direct rebound effect for personal automotive transport is likely to lie somewhere between 10% and 30%. Despite wide differences in data, methodologies and elasticity measures, most of these studies provide estimates in this range—which suggests that the findings are relatively robust. Moreover, most studies assume that the response to a change in fuel prices is equal in size to the response to a change in fuel efficiency, but opposite in sign (i.e., \( \eta_p(S) = -\eta_p(S) \)). Few studies test this assumption explicitly and those that do are either unable to reject the hypothesis that the two elasticities are equal in magnitude, or find that the

---

7 On the basis of 22 cross-sectional studies, Hanley et al. estimate a mean value for \( \eta_p(S) \) of 0.30, with a standard error of 0.22.

8 For personal automotive transport, time costs \( (P_t) \) form an important component of the total cost \( (P_t) \) of the energy service. Time costs should be correlated with average incomes \( (P_t) \); Hence, as incomes increase, the relative importance of time costs should also increase \( (P_t/P_t) \), while the relative importance of energy costs should reduce \( (P_t/P_t) \). Hence, a proportional improvement in energy efficiency should have a smaller effect on the total cost of travel for higher income groups and hence on the demand for travel by those groups.

9 Problems include the inappropriate use of total gasoline consumption as a proxy for the fuel consumed by automobiles, the uncertainties in estimates of distance driven obtained from roadside counts or surveys of vehicle users, and the uncertainties over the number of cars in use in each year (Schipper et al., 1993; Sorrell, 1992). For example, many countries measure automobile fuel consumption \( E \) as the residual after the estimated fuel consumption of trucks, buses and other vehicles is removed. This estimate is then calibrated against estimates of automobile fuel efficiency \( (e) \) and total distance driven \( (S) \), thereby creating a circularity between the estimates of these three variables. Hence, using any of these variables to estimate the third may be inappropriate. These measurement difficulties have been compounded by the increasing use of light trucks for personal transport in the US and the increasing penetration of diesel vehicles in Europe. While the difficulties can be resolved to some extent, the size of the measurement error may in some circumstances be comparable to the size of the variation in the relevant variables.

10 A meta-analysis by Espey (1998) found no significant differences between the US and Europe in the estimated long-run own-price elasticity of gasoline demand \( (\eta_p(E)) \).
fuel-efficiency elasticity is less than the fuel cost per kilometre elasticity (i.e. \( \eta_1(S) < \eta_{PE}(S) \)). The implication is that the direct rebound effect may lie towards the lower end of the above range (i.e. around 10%). Overall, it appears that a 10% increase in income reduces the short-run direct rebound effect by 0.58%. Using US average values of income, urbanisation and fuel prices over the period 1997–2001, they find a direct rebound effect of only 2.2% in the short-term and 10.7% in the long-term—approximately half the values estimated from the full data set. If this result is robust, it has some important implications. However, two-fifths of the estimated reduction in the rebound effect derives from the assumption that the magnitude of this effect depends upon the absolute level of fuel costs per kilometre. But since the relevant coefficient is not statistically significant, this claim is questionable.

Although methodologically sophisticated, the study is not without its problems. Despite covering 50 states over a period of 36 years, the data provides relatively little variation in vehicle fuel efficiency making it difficult to determine its effect separately from that of fuel prices and reducing the amount of variation in energy service costs (\( P_0 \)). Direct estimates of \( \eta_1(S) \) are small and statistically insignificant, which could be interpreted as implying that the direct rebound effect is approximately zero, but since this specification performs rather poorly overall, estimates based upon \( \eta_{PE}(S) \) are preferred. Also, the model leads to the unlikely result that the direct rebound effect is negative some states. This raises questions about Small and Van Dender’s (2005) use of the model for projecting declining rebound effects in the future, since increasing incomes could make the estimated direct rebound effect negative in many states (Harrison et al., 2005).

### Box 1—The declining direct rebound effect

Small and Van Dender (2005, 2007) provide one of the most methodologically rigorous estimates of the direct rebound effect for personal automotive transport. They estimate an econometric model explaining the amount of travel by passenger cars as a function of the cost per mile and other variables. By employing simultaneous equations for vehicle numbers, average fuel efficiency and vehicle miles travelled, they take into account of the fact that fuel efficiency is likely to be endogenous: i.e. more fuel-efficient cars may encourage more driving, while the expectation of more driving may encourage the purchase of more fuel-efficient cars. Their results show that failing to allow for this can lead the direct rebound effect to be overestimated.

Small and Van Dender use aggregate data on vehicle numbers, fuel efficiency, gasoline consumption, vehicle miles travelled and other variables for 50 US states and the District of Columbia covering the period 1961–2001. This approach provides considerably more observations than conventional aggregate time-series data, while at the same time providing more information on effects that are of interest to policymakers than do studies using household survey data. However, measurement of key variables such as distance travelled is prone to error (Schipper et al., 1993).

Small and Van Dender estimate the short-run direct rebound effect for the US as a whole to be 4.5% and the long-run effect to be 22%. The former is lower than most of the estimates in the literature, while the latter is close to the consensus. However, they estimate that a 10% increase in income reduces the short-run direct rebound effect by 0.58%. Using US average values of income, urbanisation and fuel prices over the period 1997–2001, they find a direct rebound effect of only 2.2% in the short-term and 10.7% in the long-term—approximately half the values estimated from the full data set. If this result is robust, it has some important implications. However, two-fifths of the estimated reduction in the rebound effect derives from the assumption that the magnitude of this effect depends upon the absolute level of fuel costs per kilometre. But since the relevant coefficient is not statistically significant, this claim is questionable.

Although methodologically sophisticated, the study is not without its problems. Despite covering 50 states over a period of 36 years, the data provides relatively little variation in vehicle fuel efficiency making it difficult to determine its effect separately from that of fuel prices and reducing the amount of variation in energy service costs (\( P_0 \)). Direct estimates of \( \eta_1(S) \) are small and statistically insignificant, which could be interpreted as implying that the direct rebound effect is approximately zero, but since this specification performs rather poorly overall, estimates based upon \( \eta_{PE}(S) \) are preferred. Also, the model leads to the unlikely result that the direct rebound effect is negative some states. This raises questions about Small and Van Dender’s (2005) use of the model for projecting declining rebound effects in the future, since increasing incomes could make the estimated direct rebound effect negative in many states (Harrison et al., 2005).

### 6. Estimates for household heating

The next best studied area for direct rebound effects is household heating. Table A9 summarises the main features and results of 12 quasi-experimental studies of household heating, including the reviews by Nadel (1993) and Milne and Boardman (2000). There is a larger ‘grey’ literature on this subject—including evaluations of energy-efficiency programmes by individual utilities—but this is relatively inaccessible.

Taken together, the reviewed studies suggest that standard engineering models may overestimate the energy savings from heating improvements by around one half—and potentially by more than this for low income households. However, overall percent shortfall is highly contingent on the accuracy of the engineering models, and attempts to calibrate these models to specific household conditions generally result in a lower shortfall. The studies use a wide range of variables to explain shortfall, but only initial energy consumption and the age of the home consistently influence the extent to which predicted savings are likely to be achieved.

The studies provide mean estimates of temperature take-back in the range from 0.14°C to 1.6°C, of which approximately half is estimated to be accounted for by the physical characteristics of the house and the remainder by behavioural change. This behavioural change is not trivial: depending upon insulation standards and external temperatures, a 1°C increase in internal temperature may increase the energy consumption for space heating by 10% or more. Estimates of the energy savings lost through temperature take-back range from 0% to 100%, but with a mean around 20%. Temperature take-back appears to be higher for low-income groups and for households with low internal temperatures prior to the efficiency measures. These two explanatory variables are likely to be correlated but few studies include measures of both. As pre-intervention room temperatures approach 21°C the magnitudes of temperature take-back decreases owing to saturation effects.

Overall, while shortfall may often exceed 50% (especially for low income households), temperature take-back only accounts for a portion of this shortfall and behavioural change only accounts for a portion of the temperature take-back. Temperature take-back would appear to reduce energy savings by around 20% on average, with the contribution from behavioural change being somewhat less. Which of these measures is best interpreted as the direct rebound effect is a matter of debate. Temperature take-back may be expected to decrease over time as average internal temperatures increase.

Relatively few econometric studies estimate \( \eta_{PE}(E) \), \( \eta(S) \) or \( \eta_{PE}(S) \) for household heating and even fewer investigate direct rebound effects. All of these studies rely upon detailed household survey data and exhibit considerable diversity in terms of the demographic groups covered, the definition and measurement of the relevant variables (e.g. \( S, E \) and \( \varepsilon \)), the extent to which various factors are controlled for (e.g. income, household type, demographics) and the methodologies employed. Appendix B illustrates the range of elasticity measures that could be used as a proxy for the direct rebound effect for household heating and suggests that the estimated size of the effect may be expected to vary widely depending on the particular measure that is chosen.

Tables A5 and A6 summarise and compare the results of nine econometric studies of household heating. Each of these include elasticity estimates that may be used as a proxy for the direct
rebound effect, but most of the studies do not mention the rebound effect and the relevant measures vary widely from one study to another. Overall, the studies provide estimates in the range 10–58% for the short-run direct rebound effect and 1.4–60% for the long-run effect.

As an illustration, Schwarz and Taylor (1995) use cross-sectional data from 1188 single family US households, including measurements of thermostat settings. They estimate an equation for the thermostat setting as a function of energy prices, external temperature, heated area, household income and an engineering estimate of the thermal resistance of the house. Their data allows them to estimate the demand space heating and hence to estimate \( \eta_R(T_{\text{set}}) \) (see Appendix B). Their results suggest a long-run direct rebound effect of between 1.4% and 3.4%.

Contrasting results are obtained by Hsueh and Gerner (1993), who use comparable data from 1281 single family detached households in the US, dating back to 1981. Their dataset includes comprehensive information on appliance ownership and demographic characteristics, which allows them to combine econometric and engineering models to estimate the energy use for space heating. On the basis of an estimate of \( \eta_R(E) \), the short-run direct rebound effect is estimated as 35% for electrically heated homes and 58% for gas heated homes.\(^\text{11}\)

As these examples demonstrate, the definition of the direct rebound effect is not consistent between studies and the behavioural response appears to vary widely between different households. Nevertheless, the econometric evidence broadly supports the conclusions of the quasi-experimental studies, suggesting a mean value for the direct rebound effect for household heating of around 20%.

7. Estimates for other household services

There are relatively few estimates of the direct rebound effect for other household energy services, owing largely to lack of data. Nadel (1993) reports the results of a number of evaluation studies by US utilities, which suggest direct rebound effects of 10% or less for lighting and approximately zero for water heating, with inconclusive results for refrigeration. We were not able to access these studies, which appear to be small-scale, short-term and methodologically weak. Instead, Tables A7 and A8 summarise the results of four peer-reviewed studies.

Two studies of household cooling (Table A7) provide estimates of the direct rebound effects that are comparable to those for household heating (i.e. 1–26%). Both are methodologically sophisticated, avoid endogeneity bias and suggest that the rebound effect may vary with external temperature. However, these are relatively old studies, conducted during the period of rising energy prices and using small sample sizes. Their results may not be transferable to other geographical areas, owing to differences in house types and climatological conditions. Also, both studies focus solely upon changes in equipment utilisation. To the extent that ownership of cooling technology is rapidly increasing in many countries, demand from ‘marginal consumers’ may be an important consideration, together with increases in system capacity among existing users.

The evidence for water heating is even more limited (Table A8), although Guertin et al. (2003) provide estimates in the range 34–38%, which is significantly larger than the results from quasi-experimental studies reported by Nadel (1993). A methodologically rigorous study of direct rebound effects for clothes washing (Box 2) suggests that direct rebound effects for ‘minor’ energy services should be relatively small (i.e. <5%). However, this study confines attention to households that already have automatic washing machines and, therefore, also excludes rebound effects from marginal consumers.

Table 1 summarises the results of our survey of estimates of the direct rebound effect. Despite the methodological diversity, the results for individual energy services are broadly comparable, suggesting that the evidence is relatively robust to different populations, measures, datasets and methodologies. Also, consideration of the potential sources of bias (see below) suggests that direct rebound effects are more likely to lie towards lower end of the range indicated here.

The results suggest that the mean long-run direct rebound effect for personal automotive transport, household heating and household cooling in OECD countries is likely to be 30% or less and may be expected to decline in the future as demand saturates and income increases. Both theoretical considerations and the limited empirical evidence suggest that direct rebound effects are significantly smaller for other household energy services. However, the same conclusion may not follow for energy efficiency improvements by producers or for low income households in developing countries and it would be inappropriate to draw conclusions about rebound effects ‘as a whole’ from this evidence (see Sorrell (2007) for a review).

8. Potential sources of bias

Most estimates of the direct rebound effect assume that the change in demand following a change in energy prices is equal to that following a change in energy efficiency, but opposite in sign. Most studies also assume that any change in energy efficiency derives solely from outside the model (i.e. energy efficiency is ‘exogenous’). In practice, both of these assumptions may be incorrect.

First, while changes in energy prices are generally not correlated with changes in other input costs, changes in energy efficiency may be. In particular, higher energy efficiency may only be achieved through the purchase of new equipment with higher capital costs than less efficient models. Hence, estimates of the direct rebound effect that rely primarily upon historical and/or cross-sectional variations in energy prices could overestimate the direct rebound effect, since the additional capital costs required to improve energy efficiency will not be taken into account (Henly et al., 1988).

Second, energy price elasticities tend to be higher for periods with rising prices than for those with falling prices (Dargay and Gately, 1994, 1995; Gately, 1992a, 1993; Haas and Schipper, 1998). For example, Dargay (1992) found that the reduction in UK energy demand following the price rises of the late 1970s was five times greater than the increase in demand following the price collapse of the mid-1980s. One explanation is that higher energy prices induce technological improvements in energy efficiency, which also become embodied in regulations (Grubb, 1995). Also, investment in measures such as thermal insulation is largely irreversible over the short to medium-term. But the appropriate proxy for improvements in energy efficiency is reductions in energy prices. Since many studies based upon time series data incorporate periods of rising energy prices, the estimated price elasticities may overestimate the response to falling energy prices. As a result, such studies could overestimate the direct rebound effect.

\(^\text{11}\) Greene and Greene (1998) ignore the estimate for gas-heated homes, despite the gas equation performing better as a result of a sample size that is four times larger. Also, the appropriate proxy for the direct rebound effects for space heating is \( \eta_R(E_{\text{total}}) \), which will be larger than \( \eta_R(E_{\text{real}}) \). The range of 35–58% could, therefore, be taken as a lower bound for the direct rebound effect for space heating implied by this study.

\[^{11}\] Greene and Greene (1998) ignore the estimate for gas-heated homes, despite the gas equation performing better as a result of a sample size that is four times larger. Also, the appropriate proxy for the direct rebound effects for space heating is \( \eta_R(E_{\text{total}}) \), which will be larger than \( \eta_R(E_{\text{real}}) \). The range of 35–58% could, therefore, be taken as a lower bound for the direct rebound effect for space heating implied by this study.
Table 1
Econometric estimates of the long-run direct rebound effect for household energy services in the OECD.

<table>
<thead>
<tr>
<th>End-use</th>
<th>Range of values in evidence base (%)</th>
<th>'Best guess' (%)</th>
<th>No. of studies</th>
<th>Degree of confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal automotive transport</td>
<td>3–87</td>
<td>10–30</td>
<td>17</td>
<td>High</td>
</tr>
<tr>
<td>Space heating</td>
<td>0.6–60</td>
<td>10–30</td>
<td>9</td>
<td>Medium</td>
</tr>
<tr>
<td>Space cooling</td>
<td>1–26</td>
<td>1–26</td>
<td>2</td>
<td>Low</td>
</tr>
<tr>
<td>Other consumer energy services</td>
<td>0–41</td>
<td>&lt;20</td>
<td>3</td>
<td>Low</td>
</tr>
</tbody>
</table>

Third, while improved energy efficiency may increase the demand for energy services (e.g. you could drive further after purchasing an energy-efficient car), it is also possible that the anticipated high demand for energy services may increase the demand for energy efficiency (e.g. you purchase an energy-efficient car because you expect to drive further). In these circumstances, the demand for energy services depends on the energy cost of energy services, which depends upon energy efficiency, which depends upon the demand for energy services (Small and Van Dender, 2005). Hence, the direct rebound effect would not be the only explanation for any measured correlation between energy efficiency and the demand for energy services. This so-called ‘endogeneity’ can be addressed through the use of simultaneous equation models, but these are relatively uncommon owing to their greater data requirements. If, instead, studies use a single equation without the use of appropriate estimation techniques, the resulting estimates could be biased. Several studies of direct rebound effects could be flawed for this reason.

Finally, consumers may be expected to take the full costs of energy services into account when making decisions about the consumption of those services and these include the time costs associated with producing and/or using the relevant service—for example, the time required to travel from A to B. Indeed, the increase in energy consumption in industrial societies over the past century may have been driven in part by attempts to ‘save time’ (and hence time costs) through the use of technologies that allow tasks to be completed faster at the expense of using more energy. For example, travel by private car has replaced walking, cycling and public transport; automatic washing machines have replaced washing by hand; fast food and ready meals have replaced traditional cooking and so on. While not all energy services involve such trade-offs, many important ones do (compare rail and air travel for example). Time costs may be approximated by hourly wage rates and since these have risen more rapidly than energy prices throughout the last century, there has been a strong incentive to substitute energy for time (Becker, 1965). If time costs continue to increase in importance relative to energy costs, the direct rebound effect for many energy services should become less important—since improvements in energy efficiency will have an increasingly small impact on the total cost of energy services (Binswanger, 2001). This suggests that estimates of the direct rebound effect that do not control for increases in time costs (which is correlated with increases in income) could potentially overestimate the direct rebound effect.

Box 1 shows how this could be particular relevant to direct rebound effects in transport. Similar reasoning suggests that the direct rebound effect may decline as the mean level of energy efficiency improves as energy costs should form a declining fraction of the total cost of energy services.

The consideration of time costs also points to an important but relatively unexplored issue: increasing time efficiency may lead to a parallel ‘rebound effect with respect to time’ (Binswanger, 2001; Jalas, 2002). For example, faster modes of transport may encourage longer commuting distances, with the time spent commuting remaining broadly unchanged. So in some circumstances energy consumption may be increased, first, by trading off energy efficiency for time efficiency (e.g. choosing air travel rather than rail) and second, by the rebound effects with respect to time (e.g. choosing to travel further).

9. Summary

In summary, the accurate estimation of direct rebound effects is far from straightforward. A pre-requisite is adequate data on energy consumption, energy services and/or energy efficiency...
which is only available for a subset of energy services. As a consequence, the evidence remains sparse, inconsistent and methodologically diverse, as well as being largely confined to a limited number of consumer energy services in the OECD. It is important to recognise that estimates of the direct rebound effect for these energy services provide no guide for the magnitude of such effects in other sectors, or for the rebound effect overall.

Under certain assumptions, estimates of the own-price elasticity of energy demand for an individual energy service should provide an upper bound for the direct rebound effect for that service. Since the demand for energy is generally found to be inelastic in OECD countries, the long-run direct rebound effect for most energy services should be less than 100%.

For personal automotive transport, household heating and household cooling in OECD countries, the mean value of the long-run direct rebound effect is likely to be less than 30% and may be closer to 10% for transport. Moreover, the effect is expected to decline in the future as demand saturates and income increases. Both theoretical considerations and the available empirical evidence suggest that direct rebound effects should be smaller for other consumer energy services where energy forms a small proportion of total costs. Hence, at least for OECD countries, direct rebound effects should only partially offset the energy savings from energy efficiency improvements in consumer energy services.

These conclusions are subject to a number of important qualifications, including the relatively limited time periods over which direct rebound effects have been studied and the restrictive definitions of energy services that have been employed. For example, current studies only measure the increase in distance driven for automotive transport and do not measure changes in vehicle size. They also fail to capture the economic and social transformations that cheaper energy services may induce over the very long term (Sorrell, 2007). Rebound effects for space heating and other energy services are also higher among low-income groups and most studies do not account for 'marginal consumers' acquiring services such as space cooling for the first time.

The methodological quality of many quasi-experimental studies is poor, while the estimates from many econometric studies appear vulnerable to bias. The most likely effect of the latter is to lead the direct rebound effect to be overestimated. Considerable scope exists for improving estimates of the direct rebound effect for the energy services studied here and for extending estimates to include other energy services. But this can only be achieved with better data.

Acknowledgements

This paper is based upon a comprehensive review of the evidence for rebound effects, conducted by the UK Energy Research Centre (Sorrell, 2007). An earlier version of the results is contained in Sorrell (2008). The financial support of the UK Research Councils is gratefully acknowledged. The authors are grateful for the advice and comments received from Manuel Frondel, Karsten Neuhoef, Jake Chapman, Nick Eyre, Blake Alcott, Horace Herring, Paolo Agnolucci, Jim Skea, Rob Cross, Phil Heptonstall and an anonymous reviewer. A debt is also owed to Lorna Greening and David Greene for their previous synthesis of empirical work in this area (Greening and Greene, 1998). The usual disclaimers apply.

Appendix A. Summary of empirical estimates

Tables A2–A8 in this Appendix summarise the main features and results of 31 econometric studies that may be used to estimate the direct rebound effect (RE). The studies are classified in terms of their level of aggregation, whether they are static or dynamic, the relevant country/region/time period, the type of data used, the basic model structure, the assumed functional form, the estimation technique employed and the elasticity measure that is used as a basis for estimating the direct rebound effect. The different types of elasticity measure are discussed in the main text, while the most common options for the other variables are summarised in Table A1. A full description and evaluation of these studies is provided in Sorrell and Dimitropoulos (2007b).

Table A9 summarises the main features and results of 12 quasi-experimental studies of the direct rebound effect for household heating. It provides information on the relevant country/region(s), broad approach, main results and key weaknesses. A full description and evaluation of these studies is provided in Sommerville and Sorrell (2007).

<table>
<thead>
<tr>
<th>Category</th>
<th>Choices available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of aggregation</td>
<td>Aggregate (country or regional level) Disaggregate (household level)</td>
</tr>
<tr>
<td>Static versus dynamic</td>
<td>Static (provides only single elasticity estimates) Dynamic (allows short- and long-run elasticities to be identified)</td>
</tr>
<tr>
<td>Type of data</td>
<td>Cross-section (XS) Time-series (TS) Pooled cross-section Panel</td>
</tr>
<tr>
<td>Model structure</td>
<td>Single equation Multi-equation Recursive Simultaneous equation Discrete/continuous Household production</td>
</tr>
<tr>
<td>Functional form</td>
<td>Linear Log-linear Double log Translog</td>
</tr>
<tr>
<td>Estimation technique</td>
<td>Ordinary least squares (OLS) Feasible generalised least squares (FGLS) Instrumental variables (IV) Two-stage least squares (2SLS) Three-stage least squares (3SLS) Fixed effects (FE) and random effects (RE) Error correction (ECM) Logit/probit/tobit Maximum likelihood (ML)</td>
</tr>
</tbody>
</table>

Notes: Single equation models using aggregate cross-sectional or time-series data are the most popular and easiest to estimate, but they can lead to biased estimates if one or more of the explanatory variables is endogenous and they cannot be used to separate the relative effect of changes in the number capacity and/or utilisation of equipment. Multi-equation models perform better in this regard but require correspondingly greater data requirements. Discrete/continuous and household production models are particularly demanding and are, therefore, largely confined to disaggregate studies using comprehensive household survey data. Some model structures may be used to estimate both short-term and long-term direct rebound effects, while others may only estimate one or the other. Since the direct rebound effect can vary over time and have multiple repercussions within sectors, studies using pooled cross-section or panel data may be more likely to provide robust estimates. However, such data sets are less widely available. Examples of all these categories are included in Tables A2–A8. See Sorrell and Dimitropoulos (2007b) for a full description of the meaning of these terms.
Table A2
Econometric estimates of the direct rebound effect for personal automotive transport using aggregate time-series or cross-section data.

<table>
<thead>
<tr>
<th>Author/year</th>
<th>Short-run rebound effect</th>
<th>Long-run rebound effect</th>
<th>Country</th>
<th>Data</th>
<th>Model structure</th>
<th>Functional form</th>
<th>Estimation technique</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mayo and Mathis (1988)</td>
<td>22%</td>
<td>26%</td>
<td>US</td>
<td>TS 1958–84, National</td>
<td>Recursive (S and ε)</td>
<td>Dynamic</td>
<td>3SLS</td>
<td>Long-run estimate questionable since coefficient on lagged dependent variable not significant. Correlation between CAFE standards and fuel prices makes individual effects difficult to estimate.</td>
</tr>
<tr>
<td>Gately (1992b)</td>
<td>9%</td>
<td>9%</td>
<td>US</td>
<td>TS 1966–88, National</td>
<td>Recursive (S and ε)</td>
<td>Static</td>
<td>OLS</td>
<td>S is the total distance travelled, with control for number of licensed drivers.</td>
</tr>
<tr>
<td>Greene (1992)</td>
<td>5–19% (linear) 13% (log-linear)</td>
<td>5–19% (linear) 13% (log-linear)</td>
<td>US</td>
<td>TS 1957–89, National</td>
<td>Single equation (S)</td>
<td>Static and dynamic</td>
<td>OLS</td>
<td>S is the total distance travelled. Various models tested, but accounting for serial correlation made lagged dependent variable insignificant. Hence, long- and short-run estimates identical.</td>
</tr>
<tr>
<td>Jones (1993)</td>
<td>13%</td>
<td>30%</td>
<td>US</td>
<td>TS 1957–89, National</td>
<td>Single equation (S)</td>
<td>Static and dynamic</td>
<td>OLS</td>
<td>Same data as Greene (1992) but different models. Long-run effect estimated to be twice as large as short-run.</td>
</tr>
<tr>
<td>Schimek (1996)</td>
<td>5–7%</td>
<td>21–29%</td>
<td>US</td>
<td>TS National</td>
<td>Recursive (S, ε and NO)</td>
<td>Dynamic</td>
<td>OLS</td>
<td>S is the total distance travelled. Tests showed the absence of endogeneity bias, ηε(S) found to be equal and opposite to ηε(S).</td>
</tr>
<tr>
<td>Wheaton (1982)</td>
<td>6%</td>
<td>6%</td>
<td>25 OECD countries</td>
<td>XS 1972</td>
<td>Recursive (S, NO and ε)</td>
<td>Dynamic</td>
<td>OLS</td>
<td>Estimate based on ηε(S) for the distance travelled per vehicle. Using ηε(S) gives RE of 50%.</td>
</tr>
</tbody>
</table>
Table A3
Econometric estimates of the direct rebound effect for personal automotive transport using aggregate panel data.

<table>
<thead>
<tr>
<th>Author/year</th>
<th>Short-run rebound effect</th>
<th>Long-run rebound effect</th>
<th>Country</th>
<th>Data</th>
<th>Model structure</th>
<th>Functional form</th>
<th>Estimation technique</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wirl, 1997</td>
<td>10–20%</td>
<td>27–30%</td>
<td>UK, France, Italy</td>
<td>Aggregate Panel (X-country)</td>
<td>Single equation (S)</td>
<td>Dynamic</td>
<td>OLS</td>
<td>S is the total distance travelled. Using distance travelled per vehicle gives a best guess of 20%. Careful attention to data quality. Equivalence of $\eta_1(S)$ and $\eta_2(S)$ not tested.</td>
</tr>
<tr>
<td>Johansson and Schipper (1997)</td>
<td>5–55% Best guess: 30%</td>
<td>27–30%</td>
<td>12 OECD</td>
<td>Aggregate Panel (X-country) 1973–1992</td>
<td>Recursive (S, $\varepsilon$, NO)</td>
<td>Double log</td>
<td>Various</td>
<td>S is the distance travelled per driver. $P_d$ found to influence $\varepsilon$ with hysteresis effect.</td>
</tr>
<tr>
<td>Haughton and Sarkar (1996)</td>
<td>9–16%</td>
<td>22%</td>
<td>US</td>
<td>Aggregate Panel (US states) 1972–1991</td>
<td>Simultaneous (S, $\varepsilon$)</td>
<td>Double log</td>
<td>2SLS</td>
<td>S is the distance travelled per-capita. RE estimated to decline with income. Tests suggest that $\eta_1(S) \neq \eta_2(S)$. Estimate based on $\eta_2(S)$ on preferred.</td>
</tr>
<tr>
<td>Small and Van Dender (2005)</td>
<td>4.5%</td>
<td>22%</td>
<td>US</td>
<td>Aggregate Panel (US states) 1961–2001</td>
<td>Simultaneous (S, $\varepsilon$, NO)</td>
<td>Double log</td>
<td>3SLS</td>
<td></td>
</tr>
</tbody>
</table>

Table A4
Econometric estimates of the direct rebound effect for personal automotive transport using household survey data.

<table>
<thead>
<tr>
<th>Author/year</th>
<th>Short-run rebound effect</th>
<th>Long-run rebound effect</th>
<th>Country</th>
<th>Data</th>
<th>Model structure</th>
<th>Functional form</th>
<th>Estimation technique</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goldberg (1996)</td>
<td>0%</td>
<td></td>
<td>US</td>
<td>Rotating panel 1984–1990 (CES)</td>
<td>Discrete/ continuous</td>
<td>Double log (utilisation equation)</td>
<td>Nested logit (discrete) and instrumental variables (utilisation)</td>
<td>Very detailed model, but estimates utilisation of new cars only. If endogeneity bias ignored, RE estimated to be 22%.</td>
</tr>
<tr>
<td>Puller and Greening (1999)</td>
<td>49%</td>
<td></td>
<td>US</td>
<td>Rotating panel 1980–1990 (CES)</td>
<td>Simultaneous equation (dynamic—single year)</td>
<td>Double log</td>
<td>2SLS</td>
<td>Confined to non-business travel. Find $P_d(\alpha) &lt; 0$ reflecting only short-term changes in driving habits. Partly explains high estimate of RE. Omission of vehicle age may lead to bias.</td>
</tr>
<tr>
<td>Greene et al. (1999)</td>
<td>23%</td>
<td></td>
<td>US</td>
<td>Pooled cross-section (travel survey)</td>
<td>Simultaneous equation</td>
<td>Double log</td>
<td>3SLS</td>
<td>RE estimated from $\eta_1(S)$ for households owning 1 to 5 vehicles—quoted figure is weighted average and relates solely to utilisation. Find $S$ is the distance travelled by household. RE estimated from $\eta_2(S)$—represents an upper bound since $P_d(\alpha) &lt; P_d(E)$.</td>
</tr>
<tr>
<td>West (2004)</td>
<td>87%</td>
<td></td>
<td>US</td>
<td>Cross-section (CES-1997)</td>
<td>Discrete/ continuous</td>
<td>Double Log (utilisation equation)</td>
<td>Nested logit (discrete) ad instrumental variables (utilisation)</td>
<td>RE estimated from $\eta_1(S)$, $\eta_2(E)$, and $\eta_2(S)$, $\eta_3(S)$, $\eta_3(E)$. Results insensitive to elasticity measure and estimation method.</td>
</tr>
<tr>
<td>Frondel et al. (2008)</td>
<td>56–66%</td>
<td></td>
<td>Germany</td>
<td>Panel</td>
<td>Single equation</td>
<td>Double log</td>
<td>Fixed/between/ random effects</td>
<td></td>
</tr>
</tbody>
</table>
### Table A5
Econometric estimates of the direct rebound effect for household heating using single-equation models.

<table>
<thead>
<tr>
<th>Author/year</th>
<th>Short-run rebound effect</th>
<th>Long-run rebound effect</th>
<th>Country</th>
<th>Data</th>
<th>Functional form</th>
<th>Estimation technique</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schwarz and Taylor (1995)</td>
<td>–</td>
<td>1.4–3.4%</td>
<td>US</td>
<td>Cross-section 1984–1985 SS: 1188</td>
<td>Double log</td>
<td>OLS</td>
<td>Measure of thermostat setting ($T_i$) and level of thermal insulation allows estimates of $\eta_S(T_i)$ and $\eta_S(S_{heat})$.</td>
</tr>
<tr>
<td>Haas et al. (1998)</td>
<td>–</td>
<td>15–48%</td>
<td>Austria</td>
<td>Cross-section SS: ~400</td>
<td>Double log</td>
<td>OLS</td>
<td>RE estimated from a number of sources, including $\eta_S(E_{heat})$, $\eta_S(S_{heat})$ and $\eta_S(E_{cool})$.</td>
</tr>
<tr>
<td>Guertin et al. (2003)</td>
<td>–</td>
<td>29–47%</td>
<td>Canada</td>
<td>Cross-section 1993 SS: 440 (188 gas; 252 elec.)</td>
<td>Double log</td>
<td>OLS</td>
<td>Use of frontier analysis to estimate $e_c$. RE estimated from $\eta_S(S_{heat})$ where $P_S = P_f/e_c$.</td>
</tr>
</tbody>
</table>

### Table A6
Econometric estimates of the direct rebound effect for household heating using multi-equation models.

<table>
<thead>
<tr>
<th>Author/year</th>
<th>Short-run rebound effect</th>
<th>Long-run rebound effect</th>
<th>Country</th>
<th>Data</th>
<th>Functional form</th>
<th>Estimation technique</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nesbakken (2001)</td>
<td>15–55%</td>
<td>(average 21%)</td>
<td>Norway</td>
<td>Cross-section 1990 SS: 551</td>
<td>Discrete-continuous</td>
<td>Logit (discrete) and instrumental variables (utilisation)</td>
<td>Various fuel combinations. RE estimated from $\eta_S(E_{heat})$. No control for $e$.</td>
</tr>
<tr>
<td>Klein (1987, 1988)</td>
<td>25–29%</td>
<td>–</td>
<td>US</td>
<td>Pooled cross-section: 1973–81 SS: 2157</td>
<td>Household production</td>
<td>3SLS</td>
<td>Simultaneous estimation of a cost function for $S$, a demand function for $S$ and an equation for the relative share of capital and fuel. RE estimated from $\eta_S(S_{heat})$, which in turn is estimated from $\eta_S(S_{heat})$.</td>
</tr>
</tbody>
</table>

### Table A7
Econometric estimates of the direct rebound effect for space cooling.

<table>
<thead>
<tr>
<th>Author/year</th>
<th>Short-run rebound effect</th>
<th>Long-run rebound effect</th>
<th>Country</th>
<th>Data</th>
<th>Functional form</th>
<th>Estimation technique</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hausman (1979)</td>
<td>4%</td>
<td>26.5%</td>
<td>US</td>
<td>Cross-section 1978 SS: 46</td>
<td>Discrete-continuous</td>
<td>Nested logit (discrete) and instrumental variables (utilisation)</td>
<td>Room air-conditioners individually metered. RE estimated from $\eta_S(E_{cool})$. Use of instrumental variables avoids endogeneity bias.</td>
</tr>
<tr>
<td>Dubin et al. (1986)</td>
<td>1–26%</td>
<td>–</td>
<td>US (Florida)</td>
<td>Cross-section 1981 SS: 214–396</td>
<td>Discrete-continuous</td>
<td>Nested logit (discrete) and instrumental variables (utilisation)</td>
<td>RE estimated from $\eta_S(E)$. $e$ is a composite of $e_c$ and $e_h$. Quasi-experimental design ensures $e$ is exogenous. Comprehensive data on structural characteristics allows $e$ to be estimated with an engineering model.</td>
</tr>
</tbody>
</table>
### Table A8
Econometric estimates of the direct rebound effect for other household energy services.

<table>
<thead>
<tr>
<th>Author/year</th>
<th>Short-run rebound effect</th>
<th>Long-run rebound effect</th>
<th>Country</th>
<th>Data</th>
<th>Functional form</th>
<th>Estimation technique</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guertin et al. (2003)</td>
<td>34–38% (water)</td>
<td></td>
<td>Canada</td>
<td>Cross-section</td>
<td>Double log</td>
<td>OLS</td>
<td>$c_e$ estimated using frontier analysis. RE estimated from $\eta_1(5)$ where $P_s = P_s c_e$.</td>
</tr>
<tr>
<td></td>
<td>32–49% (appliances/</td>
<td></td>
<td></td>
<td>1993 SS: 440</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>lighting)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Davis (2007)</td>
<td>&lt; 5.6 clothes washing</td>
<td></td>
<td>US</td>
<td>Panel 1997 SS: 98</td>
<td>Double log</td>
<td>Fixed effects</td>
<td>RE estimated from $\eta_1(5)$. Quasi-experimental study, so $c_e$ is exogenous.</td>
</tr>
</tbody>
</table>

### Table A9
Estimates of the direct rebound effect from quasi-experimental studies of household heating.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Location</th>
<th>Approach</th>
<th>Shortfall</th>
<th>Temperature take-back</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hirst et al. (1985)</td>
<td>USA</td>
<td>Before–after study of 79 households who received subsidies for efficiency improvements. Internal temperature changes estimated through analysis of billing data and external temperatures.</td>
<td>50% of potential savings.</td>
<td>Mean 0.4 °C, but higher 1.3 °C for low-income households. Mean loss of 11% of potential savings.</td>
<td>No control group and risk of selection bias. Take-back estimates not statistically significant (high variance between households). Acknowledged inaccuracies in method of estimating internal temperatures.</td>
</tr>
<tr>
<td>Hirst (1987)</td>
<td>USA</td>
<td>Before–after study of 210 households who received subsidies for efficiency improvements. Compared to control group of 38 eligible non-participants. Internal temperature changes estimated through analysis of billing data and external temperatures.</td>
<td>20% of potential savings.</td>
<td>Mean 0.2 °C for first year participants and 0.7 °C for second year participants. Mean loss of 5% and 25%, respectively of potential savings.</td>
<td>Selection bias. Take-back estimates not statistically significant (high variance between households). Acknowledged inaccuracies in method of estimating internal temperatures.</td>
</tr>
<tr>
<td>Hirst et al. (1989)</td>
<td>USA</td>
<td>Multiple regressions of before–after electricity bills and cross-sectional comparisons with household explanatory variables.</td>
<td></td>
<td></td>
<td>Illustrates influence of age of house on achieved savings. Unclear how savings estimates were obtained.</td>
</tr>
<tr>
<td>Dinan and Trumble (1989)</td>
<td>USA</td>
<td>Before–after monitoring of average internal temperatures of 310 households receiving subsidises for energy-efficiency improvements. Extensive collection of socio-economic and demographic data.</td>
<td></td>
<td>Mean 0.3 °C; but higher (0.4 °C) in low-income households. Mean loss of 5% of potential savings.</td>
<td>Monitored full heating year before and after efficiency improvement. Direct monitoring of internal temperature, but only in central living area. No control group. Results suggest both physical and behavioural factors contribute to take-back.</td>
</tr>
<tr>
<td>Megdal et al. (1993)</td>
<td>Southern USA</td>
<td>Estimates take-back for three household efficiency programmes using a combination of engineering evaluations and econometric techniques.</td>
<td></td>
<td>Mean 0.6–1.1 °C. Mean loss of 18–40% depending upon programme (higher for low-income groups).</td>
<td>Novel techniques and triangulation of results. But poorly described, so unclear how estimates obtained. Calibrating engineering models reduced the gap between actual and predicted savings by 10%.</td>
</tr>
<tr>
<td>Nadel (1993)</td>
<td>USA</td>
<td>Review of 9 utility evaluations of space heating-efficiency measures.</td>
<td>Median 0.14 °C. All &lt; -0.5 °C.</td>
<td></td>
<td>Original studies not accessible. Little analysis of methodologies used. Few studies used a control group.</td>
</tr>
<tr>
<td>Milne and Boardman (2000)</td>
<td>UK and USA</td>
<td>Empirical analysis of before–after temperature measurements from 13 UK efficiency projects in mostly low-income households from the 1970s and 1980s.</td>
<td>Loss of potential energy savings depends upon initial whole-house average temperature: 50% for 14 °C, 30% for 16.5 °C, 20% for 19 °C and zero for 20 °C.</td>
<td></td>
<td>Small-scale studies covering short time periods and various types of energy-efficiency improvement. No control groups and data and methodologies not transparent. Results suggest that take-back varies with nature of energy efficiency improvement and that comfort depends on factors other than temperature.</td>
</tr>
</tbody>
</table>
Insulation resulted in 10% savings, but upgrade of boiler had no effect. 65–100% of potential savings as temperature take-back. 

Cross-sectional study, but methodologically robust with large sample size. Larger take-back in dwellings with lower initial temperatures. Large shortfall due in part to poor installation of insulation and behavioural responses such as increased window in warmer properties.

Mean shortfall 55%. Mean 0.57 °C (0.49 °C for low-income groups, but not significant) 10% of potential savings as temperature take-back.

Small sample size and no control group.


1369

Appendix B. Elasticity measures for space heating

For household heating, the dependent variable for an estimate of the direct rebound effect could either be heat outputs ($S_{\text{heat}}$) or energy inputs ($E_{\text{heat}}$). The relationship between the two may be expressed as $S_{\text{heat}} = e_c E_{\text{heat}}$, where $e_c$ is the energy efficiency of the heating system.

However, the energy service provided by the heating system is thermal comfort. To a first approximation, this may be measured by the mean internal temperature ($T_i$) of the occupied areas within the duration of occupation. It may be possible to directly monitor either the mean internal temperature of the occupied areas or the thermostat setting, but these two measures will not be equivalent (Vine and Barnes, 1989). For example, daily average household temperatures will generally increase following improvements in thermal insulation, even if the heating controls remain unchanged. This is because insulation contributes to a more even distribution of warmth around the house, reduces the rate at which a house cools down when the heating is off and delays the time at which it needs to be switched back on (Milne and Boardman, 2000). Also, thermal comfort is determined not just by internal temperature, but also by air velocity, humidity, the temperature of radiant surfaces and other factors (Dewees and Wilson, 1990; Frey and Labay, 1988). Hence, if only internal temperatures or thermostat settings are monitored, the contribution of other factors to thermal comfort will be overlooked.

The amount of energy consumption ($E_{\text{heat}}$) that is required to maintain a particular internal temperature for a specified period will depend upon the heated area or volume, the difference between internal and external temperatures, occupancy levels, solar gain, the energy efficiency of the heating system ($e_c$) and the thermal resistance of the house ($e_h$). Thermal resistance, in turn, will depend upon a number of factors such as building materials, thickness of insulation, windows, air infiltration, humidity and so on. Hence, an appropriate independent variable for an estimate of the direct rebound effect could be $e_c e_h$, a combination of the two, or one of the elements that comprise $e_h$ such as the thickness of loft insulation. Improvements in $e_h$ (e.g. better insulation) may

<table>
<thead>
<tr>
<th>Authors</th>
<th>Location</th>
<th>Approach</th>
<th>Shortfall</th>
<th>Temperature take-back</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Henderson et al.</td>
<td>UK</td>
<td>Two before–after studies of participants in energy-efficiency programmes for electrically</td>
<td>Mean shortfall: First study 68%. Second study 60%.</td>
<td>0.4 °C take-back in second study, corresponding to ~18% of potential savings (~30% of shortfall).</td>
<td>No control groups and evidence of selection bias. First study found negative relationship between shortfall and energy use prior to installation, but no significant difference between income groups. Shortfall found to depend on factors other than behavioural response.</td>
</tr>
<tr>
<td>Energy Savings</td>
<td>UK</td>
<td>Before–after comparison of gas consumption data from 499 low-income, gas-heated households</td>
<td>Mean shortfall 55%.</td>
<td></td>
<td>No control group shortfall declines and savings rise in proportion to pre-installation fuel consumption. Results very sensitive to method chosen for weather correcting consumption data.</td>
</tr>
<tr>
<td>Trust (2004)</td>
<td></td>
<td>Monitoring of living room, bedroom and external temperatures in 1804 low-income households</td>
<td>1.6 °C rise in living room;</td>
<td></td>
<td>Cross-sectional study, but methodologically robust with large sample size. Wide variation in initial temperatures and larger take-back in dwellings with lower initial temperatures.</td>
</tr>
<tr>
<td>Oreszeczyn</td>
<td>UK</td>
<td>Monitoring of living room, bedroom and external temperatures in 1804 low-income households</td>
<td>2.8 °C rise in bedrooms.</td>
<td></td>
<td>Cross-sectional study, but methodologically robust with large sample size. Larger take-back in dwellings with lower initial temperatures. Large shortfall due in part to poor installation of insulation and behavioural responses such as increased window in warmer properties.</td>
</tr>
<tr>
<td>Hong et al. (2006)</td>
<td></td>
<td>Monitoring of living room and bedroom temperatures in 1372 low-income households over 2–4</td>
<td>Insulation resulted in 10% savings, but upgrade of boiler had no effect. 65–100% of potential savings as temperature take-back.</td>
<td></td>
<td>Cross-sectional study, but methodologically robust with large sample size. Larger take-back in dwellings with lower initial temperatures. Large shortfall due in part to poor installation of insulation and behavioural responses such as increased window in warmer properties.</td>
</tr>
<tr>
<td>Martin and Watson (2006)</td>
<td>UK</td>
<td>Measurement of fuel consumption and whole house internal temperatures in 59 low-income households for 12 weeks before and after insulation improvements.</td>
<td>Mean shortfall 55%.</td>
<td>Mean 0.57 °C (0.49 °C for low-income groups, but not significant) 10% of potential savings as temperature take-back</td>
<td>Small sample size and no control group.</td>
</tr>
</tbody>
</table>

Table A9 (continued)
affect any or all of the factors that determine thermal comfort while improvements in εc will generally only affect the amount of fuel required to provide a particular level of heat output (S_{thermal}). As a result, a change in εc may be expected to have a different effect on the demand for S_{thermal} than a change in εp. Also, a change in one of the factors determining εp (e.g. cavity wall insulation) may have a different effect on the demand for S_{thermal} than a change in another factor (e.g. window glazing).

The direct rebound effect may also be estimated from the energy cost elasticity of the demand for space heating (ηp(E_{space})), provided it is reasonable to assume that the change in demand for space heating following a change in energy efficiency is equivalent to that following a change in energy prices, but opposite in sign. This assumption appears more likely to hold for changes in εc rather than changes in εp, since the latter may change thermal comfort in a variety of ways independently of conscious behavioural change.

The direct rebound effect may also be estimated from the own-price elasticity of the demand for energy for space heating (ηp(E_{space})), but this measure is likely to overestimate the effect owing to the neglect of price-induced energy efficiency improvements. However, more accurate estimates of the short-run direct rebound effect may be obtained if energy efficiency can be controlled for within the specification-ηp(E_{space})-where ε refers to εp, εc or (preferably) both. This elasticity effectively captures short-run changes in equipment utilisation in response to changes in energy prices. Under certain assumptions, this may be comparable to the change in equipment utilisation in response to changes in energy efficiency.

Energy use for heating is rarely sub-metered, but may be estimated from data on total household energy use (E_{total})-heating degree days, appliance ownership and other factors. Direct rebound effects may also be estimated from ηp(E_{total}), but the accuracy of this will depend upon the proportion of fuel/electricity consumption used for space heating.

References


Nadel, S. 1993. The Take-Back Effect—Fact or Fiction. ACEEE.


Energy-capital substitution and the rebound effect

Steve Sorrell

Paper for the 7th BIEE Academic Conference

The New Energy challenge: Security and Sustainability
St. John’s College, Oxford, 24-25th September 2008

1. Sussex Energy Group, SPRU (Science & Technology Policy Research), Freeman Centre, University of Sussex, Falmer, Brighton, BN1 9QE, Tel: +44 1273 877067, Fax: 01273 685865, Email: s.r.sorrell@sussex.ac.uk
Abstract

In neoclassical production theory, the scope for substitution between two factors of production is measured by the elasticity of substitution. The sign of this measure is commonly used to define two inputs as either ‘substitutes’ or ‘complements’. But despite more than one hundred empirical studies over the last 30 years, there is no consensus on whether energy and capital may be considered and substitutes for complements. The issue may also be relevant to the size of any ‘rebound effect’ from improved energy efficiency, since it has been suggested that the greater the ease of substitution between energy and other factors of production, the larger will be the rebound effect.

This paper seeks to clarify the relationship between elasticities of substitution and the rebound effect. It first summarises the different definitions of the elasticity of substitution and highlights some conceptual and methodological difficulties in estimating their magnitude. It then shows how assumptions about elasticities of substitution underpin both theoretical and modelling studies of the rebound effect and how such studies could easily lead to flawed conclusions. It argues that: first, the assumptions used within many energy-economic models have little empirical foundation; second, the relationship between elasticities of substitution and the rebound effect is more complex than has previously been suggested; and third, large rebound effects are possible even in circumstances where energy and capital are found to be complements.

Keywords

Rebound effect; elasticity of substitution, energy-capital substitution
1 Introduction

Within neoclassical production theory, the scope for substitution between two inputs \((i,j)\), or two groups of inputs, is determined by the ‘elasticity of substitution’ \((EoS_{ij})\) between those inputs. Large or positive values of the elasticity of substitution between energy and other inputs imply that a particular sector or economy is more ‘flexible’ and should adapt relatively easily to changes in energy prices, while small or negative values suggest that increases in energy prices may have a disproportionate impact on productivity and growth (Hogan and Manne, 1970). Beginning with Berndt and Wood (1975), a large number of studies have suggested that energy and capital are ‘complements’, which implies that, holding output fixed, an increase in energy prices will reduce the rate of capital formation as well as the demand for energy (Apostolakis, 1990; Broadstock, et al., 2007; Hogan, 1979; Koetse, et al., 2008). Such a result would suggest that energy and capital are closely linked in economic production and that increases in energy prices could have significant economic impacts.

The ease of substitution between energy and other inputs may also be relevant to the size of the ‘rebound effect’ from improved energy efficiency – meaning the difference between the actual energy savings and those predicted from engineering models (Box 1). For example, Saunders (2000b) has argued that:

“….the ease with which fuel can substitute for other factors of production (such as capital and labour) has a strong influence on how much rebound will be experienced…..the greater this ease of substitution, the greater will be the rebound” (Saunders, 2000, p. 443).

Rebound effects are difficult to measure but may be estimated using Computable General Equilibrium (CGE) models of the macroeconomy. These are based upon neoclassical production theory and the assumptions made for the elasticities of substitution between energy and other inputs can have a major influence on the model results – including the estimated size of any rebound effects (Allan, et al., 2006a; b; Grepperud and Rasmussen, 2004).

This paper seeks to clarify the relationship between elasticities of substitution and the rebound effect. It first summarises the different definitions of the elasticity of substitution and highlights some conceptual and methodological difficulties in estimating their magnitude. It then shows how assumptions about elasticities of substitution underpin both theoretical and modelling studies of the rebound effect and how such studies could easily lead to flawed conclusions. It argues that: first, the assumptions used within many energy-economic models have little empirical foundation; second, the relationship between elasticities of substitution and the rebound effect is more complex than the above quote suggests; and third, large rebound effects are possible even in circumstances where energy and capital are found to be complements.
Box 1 The rebound effect

The nature, operation and importance of rebound effects are the focus of a long-running dispute within energy economics. On the micro level, the question is whether improvements in the technical efficiency of energy use can be expected to reduce energy consumption by the amount predicted by simple engineering calculations. For example, will a 20% improvement in the fuel efficiency of passenger cars lead to a corresponding 20% reduction in motor-fuel consumption for personal automotive travel? Economic theory suggests that it will not. Since energy efficiency improvements reduce the cost of energy services such as travel, the consumption of those services may be expected to increase. For example, since the cost per mile of driving is cheaper, consumers may choose to drive further and/or more often. This increased consumption of energy services may be expected to offset some of the predicted reduction in energy consumption.

This so-called direct rebound effect was first brought to the attention of energy economists by Daniel Khazzoom (1980) and has since been the focus of much research (Greening, et al., 2000; Sorrell, 2007). But even if the direct rebound effect is zero for a particular energy service (e.g. even if consumers choose not to drive any further in their fuel efficient car), there are a number of other reasons why the economy-wide reduction in energy consumption may be less than simple calculations suggest. For example, the money saved on motor-fuel consumption may be spent on other goods and services that also require energy to provide. These so-called indirect rebound effects can take a number of forms that are generally difficult to quantify (Sorrell, 2007). Both direct and indirect rebound effects apply equally to energy efficiency improvements by consumers, such as the purchase of a more fuel efficient car, and energy efficiency improvements by producers, such as the use of energy efficient motors in machine tools. The sum of the two is the overall or economy-wide rebound effect and may be expected to vary widely depending upon the nature of the energy efficiency improvement and the time-frame and system boundary under consideration.

2 Defining elasticities of substitution

The concept of substitution can be formalised through the neoclassical production function, which describes the maximum output \(Y\) obtainable from inputs of capital \((K)\), labour \((L)\), energy \((E)\) and materials \((M)\) given existing technical opportunities \((Y=f(K,L,E,M))\) (Beattie and Taylor, 1993; Chambers, 1988). Empirical studies frequently use the dual cost function which specifies the minimum cost of producing a given level of output as a function of the unit price of inputs \((C=g(Y;pl,pL,pE,pM))\).

Economists typically use a number of standard functional forms for production or cost functions, such as the Cobb-Douglas, Constant Elasticity of Substitution (CES) and the Translog (Chambers, 1988; Saunders, 2008). These are based on standard, but necessarily restrictive assumptions, such as perfectly competitive behaviour, continuity, positivity (of first and second derivatives) and concavity.

The elasticity of substitution \((EoS_{ij})\) is intended to measure the ease with which one input \((i)\) can be substituted for another \((j)\), holding output fixed. The sign of this elasticity is frequently used to classify inputs as either substitutes or complements. However, there are a number of definitions of the elasticity of substitution and whether two inputs may be described as substitutes or complements depends upon the definition being used. The lack of clarity in definitions, together with inconsistency in terminology can make the empirical literature in this area difficult to interpret (Stern, 2004).
The most common measures of the elasticity of substitution are the: Hicks/Direct Elasticity of Substitution (HES); the Cross Price Elasticity (CPE); the Allen-Uzawa Elasticity of Substitution (AES); and the Morishima Elasticity of Substitution (MES). To clarify their definitions, let: $x_i =$ level of input $i$; $p_i =$ unit price of input $i$; $f_i =$ marginal productivity of input $i$ ($\frac{\partial f}{\partial x_i}$); and $s_i =$ share of input $i$ in the value of output ($s_i = x_i p_i / p_y$). The mathematical definitions of each EoS measure are then summarised in Table 1. For further information on these definitions, see Broadstock, et al. (2007), Frondel (2004), Stern (2004) and Sato and Koizumi (1973).

The HES was introduced by Hicks (1932) as a measure of the ease with which a decrease in one input could be compensated by an increase in another while holding output constant. The definition refers to movement along a partial isoquant of a production function and is a scale-free measure of the curvature of this isoquant. The less the curvature the easier it is to substitute between the two inputs and the smaller the HES ($HES_{ij} \geq 0$). The original Hicks definition applied to a production function with only two inputs and while it can be extended to multi-input production functions, the assumption that other inputs are held fixed makes it of limited value. Hence, while the HES has an important role both theoretically and in the production functions used in energy-economic models, empirical studies tend to estimate the CPE, the AES or the MES. These are typically derived from an econometrically estimated cost function and the definitions allow all inputs to vary in response to changes in input prices. It is only with respect to these measures that factors can meaningfully be classified as substitutes and complements. In all cases, substitution between two inputs is ‘easier’ when the elasticity of substitution between them is greater.

The great majority of empirical studies use the AES and classify inputs as either substitutes or complements on the basis of the sign of this measure. This reflects the most common understanding of these terms, which is the effect of a change in the price of one input on the demand for another. However, exactly the same information is provided by the CPE - the AES just divides the CPE by the cost share of input one of the inputs, which is not necessarily very helpful since it implies that the quantitative value of the AES lacks meaning. Hence, in many circumstances, it would preferable to estimate the CPE (Frondel, 2004). The sign of the MES is less useful as a definition of substitutes or complements, since in nearly all cases the MES is positive. However, the MES (like the CPE but unlike the AES) is asymmetric, which is more representative of actual economic behaviour.

The magnitude of each measure relative to unity provides an indication of how the cost share of each input will change in either absolute or relative terms (Table 1). This is sometimes used to classify inputs as either ‘weak substitutes’ ($EoS<1$) or ‘strong substitutes’ ($EoS>1$). Finally, all EoS measures are restricted in their usefulness since they do not capture the impact of rising input prices on the level of output.
Table 3.1 Comparing different definitions of the elasticity of substitution

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
<th>Output</th>
<th>Other inputs ((x_k \neq x_i, x_j))</th>
<th>Other input prices ((p_k \neq p_i, p_j))</th>
<th>Type</th>
<th>Substitutes (complements)</th>
<th>Input shares</th>
<th>Symmetrical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hicks/Direct</td>
<td>(HES_{ij} = -\frac{\partial \ln(x_i / x_j)}{\partial \ln(f_j / f_i)})</td>
<td>Fixed</td>
<td>Fixed</td>
<td>Fixed</td>
<td>Two input, two price</td>
<td>All inputs are substitutes</td>
<td>(\frac{\partial (s_i / s_j)}{\partial (x_i / x_j)} \geq 0)</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>(HES_{ij} = \frac{\partial \ln(x_i / x_j)}{\partial \ln(p_j / p_i)})</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross price</td>
<td>(CPE_{ij} = \frac{\partial \ln x_i}{\partial \ln p_j})</td>
<td>Fixed</td>
<td>Variable</td>
<td>Fixed</td>
<td>One input, one price</td>
<td>(CPE_{ij}&gt;0 (&lt;0))</td>
<td>(\frac{\partial \ln s_j}{\partial \ln p_j} \geq 0)</td>
<td>No</td>
</tr>
<tr>
<td>Allen</td>
<td>(AES_{ij} = \frac{1}{s_j} \frac{\partial \ln x_i}{\partial \ln p_j})</td>
<td>Fixed</td>
<td>Variable</td>
<td>Fixed</td>
<td>One input, one price</td>
<td>(AES_{ij}&gt;0 (&lt;0))</td>
<td>(\frac{\partial \ln s_j}{\partial \ln p_j} \geq 0)</td>
<td>Yes</td>
</tr>
<tr>
<td>Morishima</td>
<td>(MES_{ij} = \frac{\partial \ln(x_i / x_j)}{\partial \ln(p_j)})</td>
<td>Fixed</td>
<td>Variable</td>
<td>Fixed</td>
<td>Two input, one price</td>
<td>(MES_{ij}&gt;0 (&lt;0))</td>
<td>(\frac{\partial \ln(s_i / s_j)}{\partial \ln p_j} \geq 0)</td>
<td>No</td>
</tr>
</tbody>
</table>

Source: Broadstock et al. (2007)
3 Estimating elasticities of substitution

In their classic study of US manufacturing over the period 1947-71, Berndt and Wood (1975) estimated a four-input (KLEM) translog cost function and found all input pairs to be AES substitutes apart from capital and energy. Since then, more than one hundred studies have estimated the EoS between capital and energy, but have failed to reach a consensus on whether energy and capital may be regarded as substitutes or complements (Broadstock, et al., 2007). Different studies have analysed different countries, sectors and time periods using different specifications, data sets and methods of estimation and frequently come to quite different conclusions. While this may be expected if the degree of substitutability depends upon the sector, level of aggregation and time period, or for the same sector in different countries. Moreover, while the authors cite a range of possible causes of these different results, there appears to be no consensus on either the relative importance of different causes or the likely direction of influence of each individual cause (i.e. whether a particular specification/assumption is likely to make the elasticity estimate larger or smaller) (Broadstock, et al., 2007).

If thirty years of research has failed to reach a consensus, it is worth asking whether the right questions are being asked and/or whether the appropriate concepts and tools are being applied. Although the estimation of substitution elasticities raises a host of theoretical and methodological issues, two appear particularly relevant to the rebound effect, namely the assumption of separability and the level of aggregation (Berndt and Christensen, 1973).

3.1 Separability

Many empirical studies make assumptions about the separability of production inputs, which implies that the marginal rate of technical substitution ($\frac{\partial x_i}{\partial x_j}$) between two inputs is unaffected by the level or price of the other inputs (Frondel and Schmidt, 2004; Leontief, 1947). If two inputs ($i,j$) are separable from a third ($k$), then the ease of substitution between $i$ and $k$ (as measured by the CPE, AES or MES) is equal to that between $j$ and $k$ (e.g. $AES_{ik}=AES_{jk}$).

Assumptions of separability are commonly used to justify either the omission of inputs for which data is unavailable (e.g. materials) or the grouping, or nesting, of different inputs. With nesting, the assumption is that producers engage in a two-stage decision process: first optimising the combination of inputs within each nest, and then optimising the combination of nests required to produce the final output. Two inputs may only be legitimately grouped within a nest if they are separable from inputs outside of the nest. For example, a (KL)E nesting structure requires that capital and labour are separable from energy. One of the contributions of Berndt and Wood (1975) was to show that the commonly used nest of capital and labour inputs (often referred to as ‘value-added’) was not separable from either energy or materials within their dataset.

But even when two inputs (e.g. $i$ and $j$) within a nest are separable from a third input (e.g. $k$), this does not mean that measures of the CPE, AES or MES between $i$ and $j$ are unaffected by the price of $k$ (Frondel and Schmidt, 2004). For example, even if capital and labour were separable from energy under the standard definition, the ease of substitution between capital and labour may still be affected by the price of energy. Frondel and Schmidt (2004) define a stricter condition of...
empirical dual separability, in which the value of $CPE_{ij}$ is unaffected by the price of $k$. For the value of $AES_{ij}$ to be similarly unaffected by the price of $k$, both the standard and empirical dual separability conditions must hold and the cost shares of inputs $i$ and $j$ ($s_i$ and $s_j$) must be unaffected by the price of $k$. Frondel and Schmidt (2004) observe that: “…. it difficult to imagine that this is possible in actual applications.” Hence, measures of substitution between $i$ and $j$ are likely to depend upon the price of other inputs, even when $i$ and $j$ are separable from those inputs.

This suggests that estimates of elasticities of substitution could be biased if they either: a) assume dual separability where it is not supported by the data; b) neglect the stricter conditions required for the invariance of the relevant elasticities; or c) omit measures of any input. Such situations may be common.

### 3.2 Aggregation

Measures of substitution are defined with respect to constant output and implicitly assume that this output consists of a single homogeneous product. Elasticities of substitution are then interpreted as relating to changes in the mix of inputs required to produce this product. But firms, subsectors and sectors may produce a range of different products with each requiring a different input mix. Hence, changes in input prices could lead to a change in the mix of products, as well as changes in the mix of inputs used to produce an individual product (Miller, 1986; Solow, 1987). More generally, changes in input prices could lead to a change in the relative contribution of individual sub-sectors to the output of a sector, or the relative contribution of individual sectors to the output of an economy. The scope for such changes is likely to increase with the level of aggregation for which the elasticities are estimated. For example, Solow (1987) illustrates how a sector or economy may still exhibit input substitution in the aggregate due to changes in product mix, even when the production of individual products is governed by fixed ratios.

At the same time, measures of energy-capital substitution may overestimate the possibility for substitution at a higher level of aggregation because they overlook the indirect energy consumption that is required to produce and maintain the relevant capital (or the indirect capital required to produce and deliver the relevant energy commodities). For example, energy is required to produce and install home insulation materials and energy efficient motors. This suggests that there may be less scope for substitution at the level of the macro-economy than suggested by analyses of individual sectors. More fundamentally, many neoclassical production functions violate the physical limits to substitution imposed by the second law of thermodynamics (Daly, 1997; Reynolds, 1999; Stern, 1997).

Therefore, consideration of potential changes in product/sector mix suggests that empirical studies at higher levels of aggregation may indicate a greater scope for substitution between energy and capital, while consideration of indirect energy consumption suggests the opposite. The balance between the two will depend upon the particular sectors and level of aggregation being studied. For example, the indirect energy consumption associated with the manufacture of insulation will only be relevant to estimates of energy-capital substitution if the sector that manufactures insulation is included in the estimates.
Also relevant are the distinctions between different types of capital (e.g. machinery and structures), labour (e.g. skilled and unskilled) and energy (e.g. electricity and fuel) (Apostolakis, 1990). There will be scope for substitution within these subcategories (e.g. between electricity and fuel), as well as between subcategories (e.g. machinery and fuel), but studies that use more aggregate measures may not reveal this. For example, aggregate energy and capital may be estimated to be substitutes, but capital may be a complement for electricity inputs and a substitute for fuel oil.\textsuperscript{13}

In sum, measures of substitution lack meaning unless qualified by the empirical circumstances to which they apply, including the level of aggregation. Hence, general statements regarding whether inputs such as energy and capital are either substitutes or complements are likely to be misleading.

### 4 The relevance of energy-capital substitution to modelling estimates of rebound effects

One approach to estimating the size of rebound effects is through CGE modelling of the macroeconomy (Allan, et al., 2006b; Glomsrod and Wei, 2005; Grepperud and Rasmussen, 2004). These models require assumptions about production structures, functional forms and the values of the relevant parameters - including those that determine elasticities of substitution (Bhattacharyya, 1996; Conrad, 1999). Allan et al. (2006b) review a number of CGE estimates of rebound effects and conclude that the assumptions made for the $HES$ between energy and other inputs have a significant influence on the results. As an illustration, Grepperud and Rasmussen (2004) estimated rebound effects to be higher in the Norwegian primary metals sector than in the fisheries sector, owing largely to the greater opportunities for substitution in the former (Grepperud and Rasmussen, 2004). Hence, if these results are to be robust, the estimates for substitution elasticities should be firmly based upon relevant empirical research. Unfortunately, there are considerable difficulties in using empirical studies to infer values of the $HES$ for CGE models. This is because CGE models typically:

- differ from empirical studies in the manner in which individual inputs are aggregated and in the level of sectoral aggregation;
- assume separability between different groups of inputs, while most empirical studies do not;
- use the nested $CES$ functional form, while most empirical studies use less restrictive functional forms such as the translog;
- require estimates of the elasticity of substitution between nests of inputs, while the parameters estimated by most empirical studies relate to individual pairs of inputs; and
- define production functions by means of the $HES$, while most empirical studies use cost functions to estimate the $AES$, the $CPE$ or the $MES$.

Blackorby and Russell (1981) show that the $AES$, $MES$ and $HES$ are identical if (and only if) there are only two inputs to the production function, the production function has a Cobb Douglas structure or the production function has a non-nested CES structure. But the two-input case is of
limited interest, the Cobb-Douglas structure is excessively restrictive and the non-nested CES requires all inputs to be equal substitutes— which is demonstrably false (McFadden, 1963). In order to provide greater flexibility in substitution possibilities, most CGE models use a nested CES functional form, in which pairs of inputs are combined together in a CES function which then combines with another input in a second CES function (Sato, 1967). Since the nested CES assumes that the HES between nests is constant, it constrains substitution possibilities to a greater degree than, for example the translog functional form that is standard in empirical studies. However, the nested CES remains popular for modelling, in part because it is easier to handle computationally. For example, a nested (KL)/(EM) CES production function could take the form:

\[ Y = \left[ a(bK^a + (1-b)E^a)^{\rho} + (1-a)(cL^\beta + (1-c)M^\beta)^{\rho} \right]^{\frac{1}{\rho}} \]  

(1)

Which may also be written as:

\[ Y = g[E^*, L^*] \]  

(2)

This is called a two-level nested CES because it contains CES production functions for two intermediate inputs \((E^*\) and \(L^*)\) embedded within a CES production function for aggregate output \((Y)\). The relevant CES functions are:

\[ E^* = (bK^a + (1-b)E^a)^{\frac{1}{\rho}} \]  

(3)

\[ L^* = (cL^\beta + (1-c)M^\beta)^{\frac{1}{\rho}} \]  

(4)

\[ Y = [a(E^*)^\rho + (1-a)(L^*)^\rho]^{\frac{1}{\rho}} \]  

(5)

This nesting structure rests on the assumption that the capital-energy composite \((E^*)\) is separable from the labour-materials composite \((L^*)\) - which is likely to be incorrect (Berndt and Christensen, 1973). Alternative nesting schemes (such as (KL)/(EM) or (EL)/(KM)) are widely used, but the appropriate choice between them is rarely tested empirically (Kemfert, 1998). If a distinction is made between different types of capital, labour or energy inputs (e.g. electricity and non-electricity), a multilevel CES can be formed, with more than one function nested within the original one (Chang, 1994).

With a nested CES, the AES between a pair of inputs belonging to different nests is equal to the HES between the nests (Berndt and Christensen, 1973). For example, the (KE)(LM) nesting structure implies that: \(AES_{KL} = AES_{KM} = AES_{EL} = AES_{EM} = HES_{E*L*}\). But the AES between a pair of inputs belonging to the same nest is not equal to HES between those inputs. Indeed, while two inputs within an individual nest are necessarily HES substitutes, they may at the same time be AES complements. The AES between these two inputs is only equal to the HES if the output of the nest is held constant.
Sato (1967) shows how the AES between a pair of inputs belonging to the same nest depends upon the HES within the nest, the value share of that nest and the HES between the nests. In the case of Equation 1:

$$AES_{KE} = HES_{E' L'} + \frac{1}{a} (HES_{KE} - HES_{E' L'})$$

(6)

Hence, it is possible for $AES_{KE}$ to be negative, despite $HES_{KE}$ and $HES_{E' L'}$ being positive. A necessary condition for this is that $HES_{E' L'} > HES_{KE}$ - or in other words, the scope for substitution between the capital-energy composite ($E^*$) and the labour-materials composite ($L^*$) must be greater than the scope for substitution between energy and capital in the production of $E^*$. A sufficient condition is that: $(HES_{E' L'} - HES_{KE})/a > HES_{E' L'}$, which suggests that $AES_{KE}$ is more likely to be negative when the share of energy services in the value of output ($a$) is small.

The key point, however, is that empirical estimates of the AES between two inputs do not easily translate into the HES parameters required for the nested CES production functions used in most CGE models. If the separability assumptions were valid, a particular nested CES could be parameterised if the function was estimated directly. But the majority of empirical studies estimate translog rather than nested CES functions and do not impose separability restrictions. Moreover even if separability restrictions are imposed with a translog, these do not ensure that estimates of the AES between two inputs are invariant to the price of other inputs since this requires the stricter conditions described by Frondel and Schmidt (2004). Furthermore, even if the stricter conditions hold, the implied nesting structure may not correspond to that required within a particular energy-economic model. Hence, the links between empirical estimates of elasticities of substitution and the assumptions made in CGE models appear to be tenuous at best.

Table 2 compares nesting structures and assumed values of EoS in a number of contemporary CGE models. Over half of these models exclude materials inputs, so therefore implicitly assume that these are (empirically dual) separable from the other inputs. These models further vary in terms of how they disaggregate and nest individual inputs (e.g. fuel and electricity) and how they model technical change. The basis for the assumed values for the HES between different inputs and nests of inputs is rarely made clear, sensitivity tests are uncommon and the values chosen vary widely between different models (although in all cases, the HES are assumed to be less than or equal to unity).

Thus: the assumptions of CGE models with regard to production structures and elasticities of substitution appear to be only tenuously linked to the empirical literature on this subject; while the empirical literature itself appears to be confused, contradictory and inconclusive. This suggests that the numerical results of CGE models - including the estimates of rebound effects - should be treated with great caution.
Table 2 Nesting structures and assumed elasticities of substitution in a selection of contemporary CGE models

<table>
<thead>
<tr>
<th>Authors</th>
<th>Nesting structure</th>
<th>Assumed values for HES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bosetti et al. (2006)</td>
<td>(KL)E</td>
<td>$HES_{K,E}=1.0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$HES_{K,L,E}=0.5$</td>
</tr>
<tr>
<td>Burniaux et al. (1992)</td>
<td>(KE)L</td>
<td>$HES_{K,E}=0$ or $0.8$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$HES_{K,L,E}=0$ or $1.0$</td>
</tr>
<tr>
<td>Edenhofer et al (2005)</td>
<td>KLE</td>
<td>$HES_{K,L,E}=0.4$</td>
</tr>
<tr>
<td>Gerlagh and van der Zwaan (2003)</td>
<td>(KL)E</td>
<td>$HES_{K,L,E}=1.0$</td>
</tr>
<tr>
<td>Goulder and Schneider (1999)</td>
<td>KLEM</td>
<td>$HES_{K,L,E,M}=1.0$</td>
</tr>
<tr>
<td>Kemfert (2002)</td>
<td>(KLM)E</td>
<td>$HES_{K,L,M,E}=0.5$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$HES_{K,L,E}=0.4$</td>
</tr>
<tr>
<td>Popp (2004)</td>
<td>KLE</td>
<td>$HES_{K,L,E}=1.0$</td>
</tr>
<tr>
<td>Sue Wing (2003)</td>
<td>(KL)(EM)</td>
<td>$HES_{K,L}=0.68$ to $0.94$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$HES_{E,M}=0.7$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$HES_{K,L,E,M}=0.7$</td>
</tr>
</tbody>
</table>

Source: van der Werf (2008)

5 The relevance of energy-capital substitution to theoretical studies of rebound effects

The theoretical relevance of elasticities of substitution to rebound effects has been illustrated by Harry Saunders (1992; 2000a; 2000b; 2008), who uses neoclassical production and growth theory to explore the consequences of ‘energy augmenting technical change’ (i.e. a form of technical change that solely improves the productivity of energy inputs). Given the shared theoretical assumptions, this approach raises very similar concerns to those identified above for CGE models.

Saunders (1992) follows Hogan and Manne (1970) in assuming a (KL)E structure for an aggregate CES production function for the macroeconomy. The $HES$ between capital and labour is assumed to be unity (i.e. a Cobb Douglas function) and ‘energy efficiency’ improvements are represented by a multiplier for energy augmenting technical change ($v_E(t) \geq 1.0$):

$$Y = v_N \left[ \left( K \right)^\gamma \left( L \right)^{1-\gamma} \right]^\rho + b(v_E(t)E)^\rho$$  \hspace{1cm} (7)
Saunders (1992) then examines the conditions under which energy augmenting technical change will lead to absolute reductions in energy consumption and concludes that this will only occur when the magnitude of the $HES$ between energy and the capital-labour composite ($HES_{KL,E} = 1/(1 - \rho)$) is less than unity. He argues that “....the greater this ease of substitution, the greater will be the rebound” (Saunders, 2000b). Moreover, Saunders shows that alternative nesting structures of the same function (i.e. $(KE)L$ and $(LE)K$) will always lead to rebound effects greater than unity (‘backfire’), regardless of the $HES$ between the nests.

Since capital and energy are in separate nests, the $AES$ between them should be equal to the $HES$ between the nests. This suggests that an estimate of the $AES$ between capital and energy could provide some information on the likelihood of backfire. But this result depends (amongst other things) upon the particular nesting structure that is assumed and hence on the validity of the separability assumptions. The $(KL)E$ nesting structure implies that the $AES$ between energy and capital is the same as that between energy and labour (Berndt and Christensen, 1973), but since most empirical estimates find these parameters to be substantially different, this nesting structure appears a poor representation of actual production relationships.

The implications of assuming a particular production structure may be made clearer by reconsidering Berndt and Wood’s (1979) graphical explanation of their finding of energy-capital complementarity. This is not the only explanation for their findings, nor is it accepted by all commentators (Griffin, 1981; Miller, 1986). But it is a plausible explanation that has some interesting implications for rebound effects.

Berndt and Wood assume a $(KE)(LM)$ nesting structure, since this was the only separability restriction that was supported by their data. For the purposes of exposition, their graphical example may be taken as illustrative of the behaviour of a $(KE)(LM)$ nested CES function, as defined by Equations 1-5, where $Y = g[E^*, L^*]$ , $E^* = h(K, E)$, $L^* = i(L, M)$ and $\left[HES_{E^*, L^*} - HES_{KE}\right]/a > HES_{E^*, L^*}$ . Berndt and Wood referred to the capital-energy composite as ‘utilised capital’, but for our purposes it is better to use the term ‘energy services’ ($E^* = h(K, E)$).

Although Berndt and Wood (1979) were not concerned with rebound effects, their graphical exposition provides a useful illustration of a rebound effect that exceeds unity. They assume a reduction in the cost of capital such as could be achieved through the introduction of investment tax credits or accelerated depreciation allowances for energy efficient technologies. This encourages capital to substitute for energy in the production of ‘energy services’ ($E^*$) which may be interpreted as an improvement in the energy efficiency ($\varepsilon$) of producing energy services (i.e. $\varepsilon = E^*/E$ has increased). The reduction in the cost of capital leads to a corresponding reduction in the cost of energy services which then encourages energy services to substitute for labour-materials in the production of output –with the result that the demand for energy services increases. In Berndt and Wood’s example, the corresponding increase in energy consumption more than offsets the reduction in energy consumption brought about by the energy efficiency improvement, leading to an overall increase in energy demand (holding output constant). Put another way, their example shows how a particular type of energy efficiency improvement –
represented by an increase in $E^*/E$ and stimulated by policies such as investment tax credits - could lead to backfire.

With this production structure, if $HES_{E^*L^*}$ is large compared to $HES_{KE}$ then the rebound effect from improvements in $e = E^*/E$ (holding output constant) should be large. This relates to the more general conclusion that rebound effects are large when the demand for ‘energy services’ ($E^*$ in this case) is elastic.

While this example gives some insight into the determinants of rebound effects, it appears to contradict Saunders’ conclusion that rebound effects will be small (large) when the scope for substitution between energy and other inputs is also small (large). In Berndt and Wood’s example, the scope for substitution between energy and capital appears to be small (since they are AES complements), but at the same time the rebound effect is very large (backfire). The main reason for the difference is that Saunders’ bases his conclusion on a different nesting structure which relies upon different assumptions about separability. In Saunders’ $(KL)E$ case, the key variable is the $HES$ between energy and a composite of capital and labour inputs. Given the separability assumptions, this is the same as the $HES$ and the AES between energy and capital. In contrast, in Berndt and Wood’s $(KE)(LM)$ example, the key variable is the magnitude of the $HES$ between energy and capital relative to the $HES$ between energy services and the labour-materials composite. Taken together, these two variables determine the AES between capital and energy - which is different from the $HES$ between capital and energy in the energy services nest.

Since $HES_{E^*L^*}$ is large relative to $HES_{KE}$, the reduction in capital costs stimulates a large increase in the demand for energy services, and the resulting increase in energy demand is more than sufficient to offset the reduction in demand that followed the substitution of capital for energy. But the high degree of substitution between energy services and the labour-materials composite is also what is responsible for the finding of energy-capital complementarity. Hence, assuming that this explanation is valid, it suggests that an empirical finding of energy-capital complementarity for a particular sector could be consistent with the potential for large rebound effects, or even backfire.

A second reason for the difference between Saunders’ conclusion and the Berndt and Wood example lies with the appropriate choice of independent variable for the rebound effect - namely, a change in energy efficiency. In Saunders theoretical work, the relevant independent variable is represented by the parameter $\nu_c(t)$, which defines energy augmenting technical change. But in broader discussions of the rebound effect, the independent variable is taken to be any relevant ratio of useful outputs to energy inputs. This ratio may apply to different levels of aggregation and may involve different ways of measuring both useful outputs and energy inputs (e.g. in thermodynamic, physical or economic terms) (Sorrell, 2007; Sorrell and Dimitropoulos, 2007). If the ratio of energy services to energy inputs ($E^*/E$) is taken as the appropriate independent variable, Berndt and Wood’s example can be interpreted as an illustration of backfire. The improvement in this ratio derives from the substitution of capital for energy following a reduction in the price of capital (as a result of investment subsidies), and does not involve any energy augmenting technical change ($\nu_c(t)$). However, the relative magnitude of the various elasticities of substitution may also be expected to influence the estimated response of the production function to such technical change.
Saunders’ conclusion may partly be reconciled with Berndt and Wood’s example if the former is reinterpreted as applying to the HES between energy services ($E^*$) and a composite of all other inputs. The magnitude of the rebound effect will be determined in part by the own price elasticity of energy services (holding output constant), which in a nested CES function will be determined by the HES between energy services and a composite of other inputs. However, just as Saunders’ original conclusion only applies if energy is separable from other inputs, this revised interpretation only applies if energy services are separable from other inputs. Also, the provision of energy services is likely to require dedicated capital ($K_{E^*}$), labour ($L_{E^*}$), materials ($M_{E^*}$) and energy ($E$) inputs, with additional capital ($K$), labour ($L$) and materials ($M$) being required for the provision of final output ($Y$). For example, a combination of the boiler and fuel inputs may provide steam (energy services) to drive separate production processes. This suggests a production function of the form:

$$Y = f[g(K, L, M), E^*(K_{E^*}, L_{E^*}, M_{E^*}, E)]$$ (8)

In principle an empirical test of Saunders’ proposition could be made through an estimate of the AES between energy services and a composite of other inputs under conditions where the $(KLM)E^*$ separability assumptions are supported by the data. But existing empirical studies measure energy rather than energy services and estimate the AES between energy and individual inputs. They therefore appear of little value in testing Saunders’ hypotheses.

Once separability restrictions are relaxed and more flexible functional forms are adopted, the relationship between elasticities of substitution and the size of the rebound effect is shown to be substantially more complex. In more recent work, Saunders (2008) shows that the magnitude of the rebound effect with a Translog cost function is a complex function of the AES between each pair of inputs, together with the cost share of energy - in contrast to the $(KL)E$ CES, where only the elasticity between energy and a composite of capital and labour inputs appears relevant. Moreover, when certain standard conditions are imposed upon the parameters of the Translog cost function, it is found to always lead to backfire. Saunders also shows that similar results apply to other, less common types of production function, including the Symmetrical Generalised Barnett and Gallant (Fourier). This suggests that Saunders early results for the CES may have led researchers to focus inappropriately upon one particular parameter (Jaccard and Bataille, 2000).

In sum, the relationship between elasticities of substitution and rebound effects appears to be far from straightforward. As a result, the existing empirical estimates of these elasticities would appear to tell us relatively little about the potential size of any rebound effects.

6 Summary

The main conclusions from this paper are as follows:

- There is only a tenuous link between empirical estimates of elasticities of substitution and the assumptions made in CGE models. Most CGE models use different types of production function from those estimated within empirical studies; use different definitions of the elasticity of substitution from those estimated with empirical studies;
combine inputs into nests while most empirical studies do not; assume that the scope for substitution within a nest is independent of the level or prices of other inputs; and make assumptions about the elasticity of substitution between different nests, while most empirical studies provide estimates of the elasticity of substitution between individual pairs of inputs. In addition: the process of compiling parameter values is rarely transparent; sensitivity tests are uncommon; the empirical studies frequently apply to different sectors, time periods and levels of aggregation to those represented by the model; and different models use widely different assumptions. All these observations suggest that the empirical basis for most CGE models is extremely weak and that the quantitative results – including the estimates of rebound effects – need to be treated with great caution.

- The relationship between elasticities of substitution and the magnitude of rebound effects is more complex than is generally assumed. Saunders’ early statement that “…the ease with which fuel can substitute for other factors of production (such as capital and labour) has a strong influence on how much rebound will be experienced” is potentially misleading. The relationship depends very much upon the distinction between energy and energy services, the particular choice of production function, the appropriate definition of the elasticity of substitution and the validity of the assumption of separability between energy services and other inputs. When these are taken into account, it is possible that large rebound effects may occur even when there is limited scope for substitution between energy and other inputs. As a result, the existing empirical estimates of elasticities substitution appear of little value for estimating the size of rebound effects.

These conclusions should not be taken to imply that empirical estimates of substitution elasticities have no value or that rebound effects cannot be usefully explored through either CGE modelling or theoretical investigations. Indeed, the recent studies by Koetse et al., (2008) Turner (2008) and Saunders (2008) provide some extremely valuable insights. But since each is based upon neoclassical production theory, the major limitations of this theory in describing real-world economic behaviour must be borne in mind when interpreting the results. There is a risk of the abstractions of neoclassical theory obscuring rather than eliminating the relevant phenomena and of giving too much weight to apparently precise quantitative estimates.
Acknowledgements

This paper is based upon Broadstock et al. (2007), which in turn forms part of a comprehensive review of the evidence for rebound effects, conducted by the UK Energy Research Centre (Sorrell, 2007). The financial support of the UK Research Councils is gratefully acknowledged. The author is extremely grateful for the advice and comments received from David Broadstock, Lester Hunt and John Dimitropoulos, and especially from Harry Saunders who has pioneered work in this area. The usual disclaimers apply.
References


The econometric estimation of production functions is prone to bias, so it is more common to estimate cost functions since the relevant independent variables (input prices) can usually be assumed to be exogenous. By differentiating the cost function with respect to the price of each input and then applying Shephard’s Lemma ($\frac{\partial C}{\partial p_i} = x_i$), equations for the share of each input in total costs can be derived (Berndt and Wood, 1975). These may be estimated econometrically to give the parameters in the cost equation which may then be used to derive measures of the elasticity of substitution.

The functional form describes the ‘shape’ of a production surface, representing the different combinations of inputs that may be used to produce a given level of output using existing technology. The form places restrictions on this shape, while the particular parameter values determined it precisely. Functional forms are chosen for their analytical tractability and should be considered as merely a convenient approximation to reality.

The following relationships can also be derived: a) $AES_s = \frac{CPE_s}{s}$; b) $MES_s = CPE_s - CPE_s$; c) $MES_s = s_s(AES_s - AES_s)$; d) $\frac{\partial \ln s}{\partial \ln p} = s_s(AES_s - 1)$; and e) $\frac{\partial \ln(s/s)}{\partial \ln p} = (MES_s - 1)$ (Broadstock, et al., 2007).

The $HES$ measures the ratio of the relative change in factor proportions to the relative change in the marginal rate of technical substitution. The marginal rate of technical substitution between input $i$ and input $j$ is given by the ratio of the marginal productivity of $j$ to the marginal productivity of $i$.

The extremes are: a linear production function, where $HES_s = \infty$, and a Leontief (fixed proportions) production function, where $HES_s = 0$. For a Cobb Douglas production function $HES=1$, while for a Constant Elasticity of Substitution (CES) production function, $HES$ is constant (as the name suggests) between 0 and infinity.

The generalisation of the two-input $HES$ to multi-input production functions is sometimes termed the Direct Elasticity of Substitution (Chambers, 1988), but in this paper, both will be referred to as the $HES$. Terminology is not consistent in this area, which can be a source of confusion. For example, Berndt and Wood (1979) refer to the $HES$ (as defined here) as the Direct elasticity of substitution, and refer to the $AES$ as the Hicks-Allen elasticity of substitution.

$APE = CPE_s / s$. As Chambers (1988) notes, this equation is the: “……most compelling argument for ignoring the Allen measure in applied analysis … The interesting measure is $CPE_s$- why disguise it by dividing by a cost share? This question becomes all the more pointed when the best reason for doing so is that it yields a measure that can only be interpreted intuitively in terms of $CPE_s$”.

According to Raj and Veall (1998), studies using the original Berndt and Wood (1975) data have produced 38 different estimates of $EoS_{KL}$, ranging from -3.94 to 10.84.

For example, the choice between cross-sectional and time series data; homogeneity assumptions; the specification of technical change; the choice of functional form; the measurement of capital; and so on (Broadstock, et al., 2007).

The primal (production function) condition is $\frac{\partial}{\partial x_i} \left[ \frac{C}{x_i} \right] = 0$ while the dual (cost function) condition is $\frac{\partial}{\partial p_i} \left[ \frac{C}{x_i} \right] = 0$. These two conditions are only equivalent when the production function is homoethic (i.e., the slope of an isoquant is constant for different levels of output) (Blackorby and Russell, 1976).
As an example, Miller (1986) argues that such effects are likely to lead the elasticity of substitution between capital and energy to be overestimated in studies using cross-sectional data. This is because the errors due to the excluded product mix variable are likely to be systematically correlated with input prices. For example, suppose the data set includes two regions or countries with very different levels of energy prices. One may expect the region with higher energy prices to specialise in less energy-intensive products, and vice versa. But these differences in product mix may be disguised by the chosen level of sectoral aggregation. If the sector is wrongly assumed to produce a homogeneous product, the data set will suggest a considerable scope for substituting capital for energy in the manufacture of that product and hence lead to a high estimate for the elasticity of substitution (Miller, 1986).

“…..From an ecological perspective, substituting capital and/or labour for energy shifts energy use from the sector in which it is used to sectors of the economy that produce and support capital and/or labour. In other words, substituting capital and/or labour for energy increases energy use elsewhere in the economy” (Kaufmann and Azary-Lee, 1990)

The estimated elasticity of substitution involving an aggregate is not necessarily a weighted average of the elasticities of substitution for the disaggregate inputs.

Defined as: 

\[ Y = (a_k K^{-\rho} + a_l L^{-\rho} + a_e E^{-\rho} + a_m M^{-\rho})^{-1/\rho} \]

For example, van der Werf (2006) tests different nesting structures for a KLE CES production functions for OECD manufacturing and finds that a \((KL)E\) structure is preferred.

Substituting parameter values gives: 

\[ AES_{ke} = \frac{1}{1-\alpha} \frac{1}{a} \left[ \frac{1}{1-\alpha - \rho} \right] \]

This is not necessarily the same as ‘energy saving’ technical change’. For a full discussion, see Sorrell and Dimitropoulos (2007).

Various separability assumptions may be tested by imposing restrictions upon the parameter values in the translog cost function.

Berndt and Wood (1979) show that \(HES_E\) is given by the product of the own price elasticity for energy services \((E^*)\) and the cost share of the factor whose price has changed.

Restrictions normally have to be imposed upon the parameter values in a Translog cost function to ensure that its behaviour is consistent with basic economic theory. In particular, the cost function must be concave, implying that the marginal product of each input declines with increasing use of that input. In many applications, such as CGE modelling, these conditions need to be satisfied for all input combinations, but empirically estimated cost functions sometimes violate these conditions (Diewert and Wales, 1987). Saunders (2008) finds that imposing a global concavity restriction means that the Translog production function always leads to backfire. However, Ryan and Wales (2000) show that if concavity is imposed locally at a suitably chosen reference point, the restriction may be satisfied at most all of the data points in the sample. Under these circumstances, the Translog may be able to represent different types of rebound effect for particular data sets – but only if it can be empirically verified that concavity is honoured across the domain of measurement.
Jevons’ Paradox revisited: The evidence for backfire from improved energy efficiency

Steve Sorrell

Sussex Energy Group, SPRU (Science & Technology Policy Research), Freeman Centre, University of Sussex, Falmer, Brighton BN1 9QE, UK

ARTICLE INFO
Article history:
Received 8 October 2008
Accepted 2 December 2008
Available online 19 January 2009

Keywords:
Rebound effect
Jevons’ Paradox
Khazzoom-Brookes postulate

ABSTRACT
Beginning with William Stanley Jevons in 1865, a number of authors have claimed that economically justified energy-efficiency improvements will increase rather than reduce energy consumption. ‘Jevons Paradox’ is extremely difficult to test empirically, but could have profound implications for energy and climate policy. This paper summarises and critiques the arguments and evidence that have been cited in support of Jevons’ Paradox, focusing in particular on the work of Len Brookes and Harry Saunders. It identifies some empirical and theoretical weaknesses in these arguments, highlights the questions they raise for economic orthodoxy and points to some interesting parallels between these arguments and those used by the ‘biophysical’ school of ecological economics. While the evidence in favour of ‘Jevons Paradox’ is far from conclusive, it does suggest that economy-wide rebound effects are larger than is conventionally assumed and that energy plays a more important role in driving productivity improvements and economic growth than is conventionally assumed.

1. Introduction

The view that economically justified energy-efficiency improvements will increase rather than reduce energy consumption was first put forward by the British economist, William Stanley Jevons in 1865 (Jevons, 1865). If it were true, ‘Jevons Paradox’ would have profound implications for sustainability. It would imply that encouraging energy efficiency as a means of reducing carbon emissions would not just be futile but positively counter-productive. The conventional assumptions of energy analysts, policymakers, business and lay people alike would be turned on their head, the costs of adjusting to climate change will be significantly greater than expected and the dominant strategies for achieving sustainability would be undermined. This would be all the more the case if, as seems logical, the Paradox applied to resource efficiency more generally, rather than just energy efficiency. But is the Paradox logically coherent? Do the arguments in its favour stand up to close scrutiny? What empirical evidence is available to suggest that it is correct?

A widely cited formulation of Jevons Paradox is as follows ‘with fixed real energy prices, energy-efficiency gains will increase energy consumption above what it would be without these gains’ (Saunders, 1992b). Harry Saunders termed this formulation the ‘Khazzoom–Brookes postulate’, after two contemporary economists (Len Brookes and Daniel Khazzoom) who have been closely associated with the idea. The choice of the term ‘postulate’ is revealing, since it indicates a starting assumption from which other statements are logically derived and which does not have to be either self-evident or supported by empirical evidence. This interpretation both reflects and encourages a debate that tends to be polarised, theoretical and inconclusive. This paper seeks to move beyond this by treating the above statement as a hypothesis and exploring some testable implications.

The paper summarises and critiques the arguments and evidence that have been cited in support of Jevons’ Paradox, focusing in particular on the work of Len Brookes and Harry Saunders. This work forms part of a broader literature on ‘rebound effects’ from energy efficiency improvements, which is reviewed by Sorrell (2007) and briefly introduced below. However, the arguments cited in support of Jevons’ Paradox generally do not include quantitative estimates of rebound effects. Instead, this work comprises a mix of theoretical argument, illustrative examples and ‘suggestive’ evidence from econometric analysis and economic history. It is these ‘indirect’ sources of evidence that are reviewed in this paper, together with a number of others that are not cited by the above authors but which appear relevant to...
their arguments. While most of these sources of evidence make no reference to Jevons’ Paradox they could arguably be used in support of the Paradox as well as challenging conventional wisdom in a number of areas. However, in all cases the evidence is suggestive rather than definitive.

The paper begins with an introduction to rebound effects, followed by a historical overview of the debate on Jevons’ Paradox, including the 19th century example of energy-efficiency improvements in steam engines, together with more contemporary examples of ‘general-purpose technologies’. This introduces the central theme of this paper: namely that the arguments and evidence used in support of Jevons’ Paradox are closely linked to broader questions regarding the contribution of energy to productivity improvements and economic growth.

Section 2 summarises Brookes’ arguments in favour of Jevons’ Paradox, identifies some empirical and theoretical weaknesses and examines whether more recent research supports his claims. Section 3 provides a non-technical summary of Saunders’ work, highlighting the dependence of these results on specific theoretical assumptions and the questions it raises for standard economic methodologies. Section 4 examines some relevant evidence on the contribution of energy to productivity improvements and economic growth and points to the interesting parallels between Brookes’ arguments and those of contemporary ecological economists. While the evidence remains ambiguous, the central argument is that energy—and by implication improved energy efficiency—plays a significantly more important role in economic growth than is assumed within mainstream economics. Section 5 highlights some of the implications of this finding for the economy-wide rebound effect. Section 6 concludes.

2. Rebound effects

The ‘rebound effect’ is an umbrella term for a variety of mechanisms that reduce the potential energy savings from improved energy efficiency. An example of a rebound effect would be the driver who replaces a car with a fuel-efficient model, only to take advantage of its cheaper running costs to drive further and more often.

The nature, operation and importance of rebound effects are the focus of a long-running debate within energy economics (Greening et al., 2000). On the micro level, the question is whether improvements in the technical efficiency of energy use can be expected to reduce energy consumption by the amount predicted by simple engineering calculations. Simple economic theory suggests that it will not. Since energy-efficiency improvements reduce the marginal cost of energy services such as travel, the consumption of those services may be expected to increase. This increased consumption of energy services may be expected to offset some or all of the predicted reduction in energy consumption.

This so-called direct rebound effect was first brought to the attention of energy economists by Khazzoom (1980) and has since been the focus of much research (Greening et al., 2000). But even if there is no direct rebound effect for a particular energy service (e.g. even if consumers choose not to drive any further in their fuel-efficient car), there are a number of other reasons why the economy-wide reduction in energy consumption may be less than simple calculations suggest. For example, the money saved on motor-fuel consumption may be spent on other goods and services that also require energy to provide. These so-called indirect rebound effects can take a number of forms, summarised in Box 1. Both direct and indirect rebound effects apply equally to energy-efficiency improvements by consumers and producers (Figs. 1 and 2).

The overall or economy-wide rebound effect from an energy-efficiency improvement represents the sum of these direct and indirect effects. It is normally expressed as a percentage of the expected energy savings from an energy-efficiency improvement. Hence, an economy-wide rebound effect of twenty per cent mean that twenty per cent of the potential energy savings are ‘taken back’ through one or more of the mechanisms indicated above. An economy-wide rebound effect of 100% means that the expected energy savings are entirely offset, leading to zero net savings for the economy as a whole. Backfire means that the rebound effects exceed 100%, leading to an overall increase in energy consumption—as Jevons predicted.

Rebound effects need to be defined in relation to particular time frame (e.g. short, medium or long term) and system boundary for the relevant energy consumption (e.g. household, firm, sector, national economy). For example, energy savings may be expected to be smaller for the economy as a whole than for the individual household or firm that is implementing an energy-efficiency improvement. The economy-wide effect is normally defined in relation to a national economy, but if energy-efficiency improvements lead to changes in trade patterns and international energy prices there may also be effects in other countries. Rebound effects may also be expected to increase in importance over time as markets, technology and behaviour adjusts. From the

---

**Box 1—Indirect rebound effects.**

**Embodied energy effects:** The equipment used to improve energy efficiency (e.g. thermal insulation) will itself require energy to manufacture and install and this ‘embodied’ energy consumption will offset some of the energy savings achieved.

**Re-spending effects:** Consumers may use the cost savings from energy-efficiency improvements to purchase other goods and services which themselves require energy to provide. As an extreme example, the cost savings from a more energy-efficient central heating system may be put towards an overseas holiday, leading to an increase in kerosene consumption.

**Output effects:** Producers may use the cost savings from energy-efficiency improvements to increase output, thereby increasing consumption of capital, labour and materials which themselves require energy to provide. If the energy-efficiency improvements are sector wide, they may lead to lower product prices, increased consumption of the relevant products and further increases in energy consumption. All such improvements increase the overall productivity of the economy, thereby encouraging economic growth, increased consumption of goods and services and increased energy consumption.

**Energy market effects:** Large-scale reductions in energy demand may translate into lower energy prices which will encourage energy consumption to increase. The reduction in energy prices will also increase real income, thereby encouraging investment and generating an extra stimulus to aggregate output and energy use.

**Composition effects:** Both the energy-efficiency improvements and the associated reductions in energy prices will reduce the cost of energy-intensive goods and services to a greater extent than non-energy-intensive goods and services, thereby encouraging consumer demand to shift towards the former.
perspective of climate change mitigation, what matters is the long-term effect on global energy consumption.

3. Historical perspectives

Jevons first developed his ideas with reference to coal use and steam engines. His central claim is that:

...it is wholly a confusion of ideas to suppose that the economical use of fuel is equivalent to a diminished consumption. The very contrary is the truth. ...Every improvement of the engine when effected will only accelerate anew the consumption of coal. (Jevons, 1865)

He cites the example of the Scottish iron industry, in which:

...the reduction of the consumption of coal, per ton of iron, to less than one third of its former amount, has been followed by a tenfold increase in total consumption, not to speak of the indirect effect of cheap iron in accelerating other coal consuming branches of industry. (Jevons, 1865)

According to Jevons, the early Savory engine for pumping floodwater out of coal mines ‘consumed no coal because its rate of consumption was too high.’ It was only with the subsequent improvements by Watt and others that steam engines became widespread in coal mines, facilitating greater production of lower-cost coal which in turn was used by comparable steam engines in a host of applications. One important application was to pump air into blast furnaces, thereby increasing the blast temperatures, reducing the quantity of coal needed to make iron and reducing the cost of iron (Ayres, 2002). Lower-cost iron, in turn, reduced the cost of steam engines, creating a positive feedback cycle (Fig. 3). It also contributed to the development of railways, which lowered the cost of transporting coal and iron, thereby increasing demand for both.

Jevons highlighted the fact that improvements in the thermodynamic efficiency of steam engines were intertwined with broader technical changes, including: ‘...contrivances, such as the crank, the governor, and the minor mechanism of an engine, necessary for regulating, transmitting, or modifying its power’ (Alcott, 2005; Jevons, 1865). These developments were essential to the increased use of steam engines as a source of motive power and demonstrate how energy-efficiency improvements are frequently linked to broader improvements in technology and overall, or ‘total factor’ productivity.2

---

2 The economic productivity (or efficiency) of a ‘factor’ input such as energy is given by the ratio of output to input for that factor. Total factor productivity (TFP) is normally defined as the rate of growth of economic output minus the weighted sum of the rate of growth of inputs— with each input being weighted by its share in the value of output (Sorrell and Dimitropoulos, 2007). Unlike changes in individual factor productivity, improvement in total factor productivity are always desirable, since they indicate that more output is being obtained from the same quantity of inputs. Standard ‘growth accounting’ techniques estimate total factor productivity as the residual growth in output that is not explained by the growth of inputs. Such improvements are frequently attributed to ‘technical change’.
Rosenberg (1989) has cited the comparable example of the Bessemer process for steel-making:

[the Bessemer process] was one of the most fuel saving innovations in the history of metallurgy [but] made it possible to employ steel in a wide variety of uses that were not feasible before Bessemer, bringing with it large increases in demand. As a result, although the process sharply reduced fuel requirements per unit of output, its ultimate effect was to increase... the demand for fuel. (Rosenberg, 1989)

The low-cost Bessemer steel initially found a large market in the production of steel rails, thereby facilitating the growth of the rail industry, and later in a much wider range of applications including automobiles. However, the mild steel produced by the Bessemer process is a very different product to wrought iron (which has a high carbon content) and is suitable for a much wider range of applications. Hence, once again, the improvements in the energy efficiency of production processes are deeply entwined with broader developments in process and product technology. While improved thermodynamic efficiency may form part of-or even a precondition for-such innovations, it does not follow that all the subsequent increase in energy consumption can be attributed to that efficiency improvement. More generally, it is possible to measure energy efficiency in a variety of ways for a variety of system boundaries (Box 2) and there is no consensus on the most appropriate definition for the purpose of estimating rebound effects.

The above examples relate to energy-efficiency improvements in the early stages of development of energy-intensive process technologies, producing goods that have the potential for widespread use in multiple applications. It is possible that the same consequences may not follow for energy-efficiency improvements in mature and/or non-energy-intensive process technologies, producing goods that have a relatively narrow range of applications. Similarly, the same consequences may not follow from

---

**Box 2—Defining energy efficiency.**

Energy efficiency may be defined as the ratio of useful outputs to energy inputs for a system. The system in question may be an individual energy conversion device (e.g. a boiler), a building, an industrial process, a firm, a sector or an entire economy. In all cases, the measure of energy efficiency will depend upon how ‘useful’ is defined and how inputs and outputs are measured (Patterson, 1996). The options include:

- **Thermodynamic measures:** where the outputs are defined in terms of either heat content or the capacity to perform useful work;
- **Physical measures:** where the outputs are defined in physical terms, such as vehicle kilometres or tonnes of steel; or
- **Economic measures:** where the outputs (and sometimes also the inputs) are defined in economic terms, such as value-added or GDP.

When outputs are measured in thermodynamic or physical terms, the term energy efficiency tends to be used, but when outputs are measured in economic terms it is more common to use the term ‘energy productivity’. The inverse of both measures is termed ‘energy intensity’. The choice of measures for inputs and outputs, the appropriate system boundaries and the timeframe under consideration can vary widely from one study to another. However, physical and economic measures of energy efficiency tend to be influenced by a greater range of variables than thermodynamic measures, as do measures appropriate to wider system boundaries. Hence, the indicator that is furthest from a thermodynamic measure of energy efficiency is the ratio of GDP to total primary energy consumption within a national economy.

Economists are primarily interested in energy-efficiency improvements that are consistent with the best use of all economic resources. These are conventionally divided into two categories: those that are associated with improvements in overall, or ‘total factor’ productivity (‘technical change’), and those that are not (‘substitution’). The latter is assumed to be induced by changes in the price of energy relative to other inputs. The consequences of technical change are of particular interest, since this contributes to the growth in economic output. However, distinguishing empirically between these two categories can be challenging, not least because changes in relative prices also induce technical change.
improvements in consumer technologies that supply energy services with a low own-price elasticity and where energy represents only a small share of total costs.

A historical perspective on rebound effects is provided by Fouquet and Pearson (2006), who present some remarkable data on the price and consumption of lighting services in the UK over a period of seven centuries (Table 1). Per capita consumption of lighting services grew much faster than per capita GDP throughout this period, owing in part to continuing reductions in the price per lumen hour. This, in turn, derived from continuing improvements in the energy efficiency of lighting technology, in combination with reductions in the real price of lighting fuel (itself, partly a consequence of improvements in the thermodynamic efficiency of energy supply). In this case, improvements in lighting technology were substantially more important than improvements in energy supply—in the ratio of 180 to 1 over the period 1800–2000.

Per capita lighting consumption increased by a factor of 6566 between 1800 and 2000, largely as a consequence of the falling cost of lighting services relative to income, but also as a result of the boost to per capita GDP provided by the technical improvements in lighting technology. Since lighting efficiency improved by a factor of one thousand, the data suggest that per capita energy consumption for lighting increased by a factor of six. In principle, the direct rebound effect could be estimated from the own-price elasticity of lighting services over this period. But this would be a questionable exercise over such a time interval, given the co-evolution and interdependence of the relevant variables. To the extent that the demand for lighting is approaching saturation in many OECD countries, future improvements in lighting efficiency may be associated with smaller rebound effects. Nevertheless, this historical perspective gives cause for concern over the potential of technologies such as compact fluorescent lamps to reduce energy consumption in developing countries.

4. Energy and economic growth

Time-series data such as that presented in Table 1 are difficult to obtain, which partly explains why relatively little research has investigated the causal links between improvements in various measures of energy efficiency and more aggregate measures of economic output and energy consumption. While many studies demonstrate strong correlations between economic output and energy consumption, the extent to which the growth in economic output can be considered a cause of the increased energy consumption, or vice versa, remains unclear. It seems likely that there is a synergistic relationship between the two, with each causing the other as part of a positive feedback mechanism (Ayres and Warr, 2002b). Hence, to explore Jevons Paradox further, it seems necessary to investigate the nature, mechanisms and determinants of economic growth—a notoriously difficult topic.

The conventional wisdom (as represented by both neoclassical and 'endogenous' growth theory) is that increases in energy inputs play a relatively minor role in economic growth, largely because energy accounts for a relatively small share of total costs (Barro and Sala-i-Martin, 1995; Denison, 1962; Gullickson and Harper, 1987; Jones, 2001). Economic growth is assumed to result instead from the combination of increased capital and labour inputs, changes in the quality of those inputs (e.g. better educated workers) and increases in total factor productivity that are frequently referred to as ‘technical change’.

This view has been contested by ecological economists, who argue instead that the increased availability of ‘high quality’ energy inputs has been the primary driver of economic growth over the last two centuries (Beaudreau, 1998, 2005; Cleveland et al., 1984; Hall et al., 1986; Kimmel et al., 2000, 1985). These authors emphasise that energy carriers differ both in their capacity to perform useful work (captured by the thermodynamic concept of ‘exergy’) and in their relative economic productivity—reflected by differences in price per kWh (Kaufmann, 1994). So for example, electricity represents a ‘higher quality’ form of energy than coal. In general, when the ‘quality’ of energy inputs are accounted for, aggregate measures of energy efficiency are found to be improving more slowly than is commonly supposed (Cleveland et al., 2000; Hong, 1983; Zarnikau, 1999).

Cleveland et al. (1984) claim that a strong link exists between quality adjusted energy use and economic output and this link will continue to exist, both temporally and cross-sectionally. This contrasts with the conventional wisdom that energy consumption has been ‘decoupled’ from economic growth. They also claim that a large component of increased labour productivity over the past 70 years has resulted from empowering workers with increasing quantities of energy, both directly and indirectly as embodied in capital equipment and technology (Cleveland et al., 1984). This contrasts with the conventional wisdom that productivity improvements have resulted from technical change. Other ecological economists argue that the productivity of energy inputs is substantially greater than the share of energy in total costs (Ayres, 2001; Ayres and Warr, 2005b)—again in contradiction to the conventional wisdom.

The conventional and ecological perspectives reflect differing assumptions and are supported by conflicting empirical evidence. A difficulty with both is that they confine attention to the relationship between energy consumption and economic growth. But the reason that energy is economically significant is that it is used to perform useful work—either in the form of mechanical work (including electricity generation) or in the production of heat (Ayres and Warr, 2005b). More useful work can be obtained with the same, or less, energy consumption through improved

---

**Table 1**

Seven centuries of lighting in the UK.

<table>
<thead>
<tr>
<th>Year</th>
<th>Price of lighting fuel</th>
<th>Lighting efficiency</th>
<th>Price of lighting services</th>
<th>Consumption of light per capita</th>
<th>Total consumption of light</th>
<th>Real GDP per capita</th>
</tr>
</thead>
<tbody>
<tr>
<td>1300</td>
<td>1.50</td>
<td>0.50</td>
<td>3.0</td>
<td>–</td>
<td>–</td>
<td>0.25</td>
</tr>
<tr>
<td>1700</td>
<td>1.50</td>
<td>0.75</td>
<td>2.0</td>
<td>0.17</td>
<td>0.1</td>
<td>0.75</td>
</tr>
<tr>
<td>1750</td>
<td>1.65</td>
<td>0.79</td>
<td>2.1</td>
<td>0.22</td>
<td>0.15</td>
<td>0.83</td>
</tr>
<tr>
<td>1800</td>
<td>1.0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1850</td>
<td>0.40</td>
<td>4.4</td>
<td>0.27</td>
<td>3.9</td>
<td>7</td>
<td>1.17</td>
</tr>
<tr>
<td>1900</td>
<td>0.26</td>
<td>14.5</td>
<td>0.042</td>
<td>84.7</td>
<td>220</td>
<td>2.9</td>
</tr>
<tr>
<td>1950</td>
<td>0.40</td>
<td>340</td>
<td>0.002</td>
<td>1528</td>
<td>5000</td>
<td>3.92</td>
</tr>
<tr>
<td>2000</td>
<td>0.18</td>
<td>1000</td>
<td>0.0003</td>
<td>6566</td>
<td>25630</td>
<td>15</td>
</tr>
</tbody>
</table>

Note: 1800 = 1.0 for all indices.

thermodynamic efficiency. Hence, if increases in energy inputs contribute disproportionately to total factor productivity improvements and economic growth, then improvements in thermodynamic efficiency may do the same. Conversely, if increases in energy inputs contribute little to productivity improvements and economic growth, then neither should improvements in thermodynamic efficiency.

5. The contribution of Len Brookes

Despite their far-reaching implications, Jevons’ ideas were neglected until comparatively recently and contemporary advocates of energy efficiency are frequently unaware of them. While the paper of Khazzoom (1980) stimulated much research and debate on direct rebound effects (Besen and Johnson, 1982; Einhorn, 1982; Greene, 1992; Greening et al., 2000; Henly et al., 1987, 1988; Lovins, 1988), most researchers ignored the long-term, macroeconomic implications that were Jevons’ primary concern. However, Jevons’ arguments have been taken up with some vigour by the British economist, Len Brookes, who has developed coherent arguments in favour of Jevons’ Paradox and combined these with critiques of government energy efficiency policy (Brookes 1990a, b, 2004, 1978, 1984, 2000). Brookes’ work has prompted a fierce response from critics (Grubb, 1990, 1992; Herring and Elliot, 1990; Toke, 1990), to which Brookes has provided a number of robust responses (Brookes, 1992, 1993).

Brookes (2000) argues that ‘The claims of what might be called the Jevons school are susceptible only to suggestive empirical support’, since estimating the macroeconomic consequences of individual improvements in energy efficiency is practically impossible. He therefore relies largely on theoretical arguments, supported by indirect sources of evidence, such as historical correlations between various measures of energy efficiency, total factor productivity, economic output and energy consumption (Schurr 1984, 1985). A key argument runs as follows:

...it has been claimed since the time of Jevons (1865) that the market for a more productive fuel is greater than for less productive fuel, or alternatively that for a resource to find itself in a world of more efficient use is for it to enjoy a reduction in its implicit price with the obvious implications for demand.

However, Brookes’ use of the term ‘implicit price’ is confusing. Individual energy-efficiency improvements do not change the price of input energy, but instead lower the effective price of output energy, or useful work. For example, motor-fuel prices may be unchanged following an improvement in vehicle fuel efficiency, but the price per vehicle kilometre is reduced. The ‘obvious implications’ therefore relate to the demand for useful work, and not to the demand for energy commodities themselves. While the former may be expected to increase, energy demand may either increase or decrease depending upon the price elasticity of demand for useful work and the associated indirect rebound effects (Sorrell, 2007).

Of course, the combined impact of multiple energy-efficiency improvements could lower energy demand sufficiently to reduce energy prices and thereby stimulate a corresponding increase in economy-wide energy demand. This forms one component of the economy-wide rebound effect. But while it is obvious that the overall reduction in energy consumption will be less than microeconomic analysis suggests, this theoretical argument appears to be an insufficient basis for claiming that backfire is inevitable.

Brookes also criticises the assumption that the demand for useful work will remain fixed while its marginal cost falls under the influence of raised energy efficiency, and the related assumption that individual energy savings can be added together to produce an estimate of what can be saved over the economy as a whole. In both cases, Brookes is highlighting the persistent neglect of both direct and indirect rebound effects in the conventional assessment of energy-efficiency opportunities. However, arguing that the economy-wide rebound effect is greater than zero is different from arguing that it is greater than one—as Jevons’ Paradox suggests.

Brookes marshals a number of other arguments in support of Jevons’ Paradox that appear more amenable to empirical test. In doing so, he highlights some important issues regarding the relationship between energy consumption, economic productivity and economic growth. The three most important arguments may be characterised as follows:

- **The ‘productivity’ argument**: the increased use of higher-quality forms of energy (especially electricity) has encouraged technical change, substantially improved total factor productivity and driven economic growth. Despite the substitution of energy for other inputs, this technical change has stimulated a sufficiently rapid growth in economic output that aggregate energy efficiency has improved at the same time as aggregate energy consumption has increased. Brookes cites two separate, but related sources of empirical evidence in support of this argument. The primary source is the work of Sam Schurr and colleagues on the historical importance of changes in energy quality (notably electrification) in driving US productivity growth (Box 3). The second, more indirect source of evidence is the work of Dale Jorgenson and others on the historical direction of technical change. Contrary to standard assumptions, Jorgenson’s results suggest that, at the level of individual sectors, technical change has been ‘energy-using’, meaning that it has increased energy intensity over time rather than reduced it.3 This work is also cited as suggestive evidence for Jevons’ Paradox by Saunders (1992b).

- **The ‘accommodation’ argument**: energy-efficiency improvements are claimed to ‘accommodate’ an energy price shock so that the energy supply/demand balance is struck at a higher level than if energy efficiency had remained unchanged (Brookes, 1984). While not immediately obvious, this argument appears to rest in part on the assumption that the per-capita income elasticity of ‘useful’ energy demand falls steadily as an economy develops, but is always greater than unity (Brookes, 1972). ‘Useful’ energy consumption is a quality-adjusted measure of per-capita energy consumption in which different energy types are weighted by their relative economic productivities (Adams and Miovic, 1968).

- **The ‘endogeneity’ argument**: a common approach to quantifying the ‘energy savings’ from energy-efficiency improvements is to hold energy intensity fixed at some historic value and to estimate what consumption ‘would have been’ in the absence of those improvements (Geller et al., 2006). The energy savings

---

3 Jorgenson and Fraumeni (1981) used econometric techniques to investigate the impact of the energy price rises of the 1970s on US manufacturing productivity. They employed the conventional neoclassical distinction between price-induced substitution between factor inputs and autonomous (i.e. non-price induced) technical change. They estimated the rate of change in the share of energy costs in the value of output of US manufacturing sectors (holding input prices constant) and found that, in 29 out of 35 sectors, the share of energy costs increased over time. This is termed ‘energy-using’ technical change. Generally, we would expect energy using technical change to be associated with increasing energy intensity. However, this may not always be the case, since it also depends upon the rate of change in total factor productivity (Sanstad et al., 2006). With energy-using technical change an increase in the price of energy will lower total factor productivity.
Box 3—Sam Schurr and the rebound effect.

Schurr (1982, 1983, 1984, 1985) and colleagues (Schurr et al., 1960) explored trends in US energy consumption, energy productivity and total factor productivity throughout the 20th century. Energy productivity was defined as the ratio of GDP to total primary energy consumption, with energy being measured on the basis of heat content. Over the period 1920–1953, energy, labour and total factor productivity were all found to be growing, while during the period 1953–1969, energy productivity was relatively unchanged while total factor productivity continued to grow rapidly. Both periods exhibited falling energy prices relative to other inputs and large increases in energy consumption, and were characterised by a decreasing share of coal in final energy consumption and an increasing share of oil and electricity. Also, in both periods, total factor productivity grew significantly faster than energy productivity.

Structural change in the economy and improvements in thermodynamic efficiency provided only a partial explanation of these trends. Since energy prices were falling in relative terms, energy substituted for other factors of production, thereby reducing energy productivity and improving capital and labour productivity. But these substitution effects were more than outweighed by technological improvements, facilitated by the availability of high-quality energy sources, which greatly improved the overall productive efficiency of the US economy. This meant that economic output increased much faster than energy consumption, owing to the greater productivity of capital and labour. The net result was to produce falling energy intensity (as measured by the energy/GDP ratio) alongside rising total energy consumption—as Jevons’ Paradox predicts.

Schurr argued that the technological improvements which drove output growth depended crucially upon the increased availability of more ‘flexible’ forms of energy (oil and electricity) at relatively low costs. These contributed to changes in industrial processes, consumer products and methods of industrial organisation that were quite revolutionary—for example, in transforming the sequence, layout and efficiency of industrial production (Schurr, 1982). Schurr’s pioneering contribution, therefore, was to highlight the importance of energy quality for productivity growth.

Brookes’ uses these observations to support his case for backfire. His argument appears to be that (a) most improvements in energy productivity are associated with proportionally greater improvements in total factor productivity; (b) improvements in total factor productivity increase economic output, leading to a corresponding increase in demand for inputs; and (c) the resulting increase in demand for energy inputs more than offsets the reduced demand for energy per unit of output. Hence, energy consumption increases while aggregate energy intensity falls.

Box 4—Endogeneity and the rebound effect.

Trends in aggregate quantities may be expressed as the product of a number of variables. For example, economy-wide energy consumption (E) may be expressed into the product of population (P), GDP per capita (A = Y/P) and energy use per unit of GDP (T = E/Y); E = PAT. Decomposition analysis allows the change in energy use over a particular period to be estimated as the sum of the change in each of the right-hand side variables. The ‘energy saved’ by energy-efficiency improvements over a particular period can then be estimated by comparing current energy consumption with an estimate of what energy consumption ‘would have been’ had energy intensity (T) remained unchanged. For example, the IEA analysed data from 11 OECD countries over the period 1973–1998 to suggest that energy use would have been 50% higher in 1998 if end-use intensity had remained at its 1973 level (Geller et al., 2006).

(Note: Strictly, this argument applies only to the use of Laspeyres indices in decomposition analysis, and not to competing approaches such as Divisia indices (Ang, 1999)). But this approach is only valid if right-hand side variables are independent of one another—or at least if any dependence is sufficiently small that it can be neglected. In contrast, Brookes argues that improved energy efficiency enables both higher affluence (A = f(T)) and higher population (P = g(T)); “… it is inconceivable that populations of today could be maintained with the technology of 500 years ago… inanimate energy allied to man’s ingenuity is what has permitted the very large increase in output in the last 200 years without which the increase in population would not have occurred. Would this increase (and the associated increase in energy consumption) have occurred if conversion efficiencies had stayed at the abysmally low levels prevailing in the early years of the nineteenth century?” (Brookes, 2000)

To capture this interdependence, the relationship could be better expressed as a system of simultaneous equations (Alcott, 2006):

\[
\begin{align*}
E &= f(P, A, T; X_E) \\
P &= g(E, A, T; X_P) \\
A &= h(E, P, T; X_A) \\
T &= i(E, P, A; X_T)
\end{align*}
\]

Hence, while a reduction in the economy-wide energy/GDP ratio (T) may have a direct effect on energy consumption through the first of these equations, it may also encourage economic growth (A), which in turn will increase the total demand for energy (E). Over the long term, rising affluence may encourage higher population levels (P), which in turn will increase energy consumption (E). Each of these changes may in turn influence the energy/GDP ratio (T). Hence, a change in one variable is likely to trigger a complex set of adjustments and the final change in energy consumption could be greater or less than the direct change. Under these conditions, decomposition analysis could overestimate (or underestimate) the energy savings from improved energy efficiency.

from energy-efficiency improvements are then taken to be the difference between the actual demand and the counterfactual scenario. But if the energy-efficiency improvements are a necessary condition for the growth in economic output, the construction of a counterfactual in this way is misconceived. This argument is not developed in detail by Brookes, but does raise questions over the use of ‘decomposition analysis’ to explore the rebound effect (Box 4).

Sorrell and Dimitropoulos (2007) describe the historical research that forms the basis for these arguments, summarise how Brookes uses this research to support his case and examine in
detail whether more recent research confirms or contradicts Brookes’ claims. They highlight a number of potential weaknesses, including the following:

- Schurr’s work applies primarily to the causal effect of shifts to higher quality energy carriers (notably electricity), rather than improvements in thermodynamic conversion efficiency or other factors that affect aggregate measures of energy efficiency. The effect of the latter on total factor productivity may not be the same as the effect of the former. Also, the patterns Schurr uncovered may not be as ‘normal’ as Brookes suggests, since the link between energy productivity and total factor productivity appears to vary greatly, both over time and between different countries and sectors.

- Neither Jorgenson and colleagues or comparable econometric studies consistently find technical change to be ‘energy-using’. Instead, the empirical results vary widely between different sectors, countries and time periods and are sensitive to minor changes in econometric specification (Norsworthy et al., 1979; Roy 2000; Sanstad et al., 2006; Welsch and Ochsen, 2005). Jorgenson’s results rest on the erroneous assumption that the rate and direction of technical change is fixed, and more sophisticated models suggest that the magnitude and sign of technical change varies between sectors and types of capital as well as over time (Sue Wing, 2008; Sue Wing and Eckaus, 2007). Moreover, even if energy-using technical change were to be consistently found, the relationship between this finding and Jevons’ Paradox remains unclear.4

- The ‘accommodation’ argument has its origins in a highly simplified theoretical model of the world economy (Brookes, 1984), which is both unconventional in approach and difficult to interpret and calibrate. The model appears to rest in part on the assumption that the per-capita income elasticity of ‘useful’ energy demand declines asymptotically to unity as income increases, thereby allowing economic output to be represented as a linear function of useful energy inputs. While an earlier study by Brookes (1972) provides some support for this hypothesis, this has not been tested by more recent studies of income elasticity of energy demand since these typically measure energy consumption on the basis of heat content (Richmond and Kauffmann, 2006a,b; Stern, 2004b).

- The ‘endogeneity’ argument is rhetorically persuasive but lacks a firm empirical basis. The relative importance of energy-efficiency improvements (however defined) compared to other forms of technical change in encouraging economic growth remains to be established.

In sum, each of these sources of evidence has empirical and theoretical weaknesses and the extent to which they (individually and collectively) support Jevons’ Paradox is open to question. Hence, while Brookes has highlighted some important issues and pointed to sources of evidence that challenge conventional wisdom, he has not provided a convincing case in support of Jevons’ Paradox.

Perhaps the most important insight from Brookes’ work is that improvements in energy productivity are frequently associated with proportionally greater improvements in overall or total factor productivity. While Schurr’s work provides evidence for this at the level of the national economy, numerous examples from the energy efficiency literature provide comparable evidence at the level of individual sectors and technologies (Pye and McKane, 1998; Sorrell et al., 2004; Worrell et al., 2003). Such examples are frequently used by authors such as Lovins (1997) to support the business case for energy efficiency. But if energy efficient technologies boost total factor productivity and thereby save more than energy costs alone, the argument that rebound effects must be small because the share of energy in total costs is small is undermined.5 Much the same applies to the contribution of improved energy efficiency to overall productivity improvements and economic growth. But this leaves open the question of whether energy-efficiency improvements (however defined) are necessarily associated with proportionally greater improvements in total factor productivity, or whether (as seems more likely) this is contingent upon particular technologies and circumstances.

6. The contribution of Harry Saunders

Harry Saunders has shown how Jevons’ Paradox is broadly supported by neoclassical production and growth theory. His work is theoretical and is necessarily based on highly restrictive assumptions. But Saunders does not claim that his work proves Jevons’ Paradox; instead, it simply provides suggestive evidence in its favour, given certain standard assumptions about how the economy operates.

Saunders (1992a,b) uses neoclassical growth theory to argue that backfire is a likely outcome of ‘pure’ energy-efficiency improvements—that is, a form of technical change (see Box 2) that improves energy productivity while not affecting the productivity of other inputs. In other words, this result does not rely on the contribution of energy to raising capital and labour productivity that is emphasised by Brookes. Neoclassical growth theory also predicts that ‘pure’ improvements in capital, labour or materials productivity will increase overall energy consumption. Since technical change typically improves the productivity of several inputs simultaneously, these models suggest that most forms of technical change will increase overall energy consumption compared to a scenario in which such improvements are not made.

Saunders’ use of the neoclassical growth model was challenged by Howarth (1997), who argued that the failure to distinguish between energy and energy services led to the probability of backfire being overestimated. However, Saunders (2000) subsequently demonstrated that backfire is still predicted by neoclassical theory when an alternative choice is made for the production function used to provide energy services. In a more recent contribution, Saunders (2008) focuses on the potential of different types of production function to generate backfire. Unlike Saunders...

---

4 The relevance of this work to the rebound effect is not made clear by either Brookes or Saunders. The primary implication is that technical change has frequently reduced energy efficiency and thereby increased overall energy consumption, even while other factors (such as structural change) have acted to decrease energy consumption. Not only is this the opposite to what is conventionally assumed in energy-economic models, it is also opposite to what is required for an empirical estimate of the rebound effect. At the same time, technical change has clearly improved the thermodynamic conversion efficiency of individual devices, such as motors and boilers. What Jorgenson and Fraumeni’s work suggests, therefore, is that improvements in thermodynamic efficiency have not necessarily translated into improvements in more aggregate measures of energy intensity at the level of industrial sectors. Similarly, a more recent study by Sue Wing and Eckaus (2007) suggests that this has not necessarily translated into improvements in more aggregate measures of energy intensity for particular types of capital (e.g. machinery). In other words, improvements in energy efficiency at one level of aggregation may have contributed to greater energy consumption at a higher level of aggregation. Hence, the relevance of these results may hinge in part upon the appropriate choice of independent variable for the rebound effect.

5 This also implies that so-called ‘win-win’ technologies may be associated with the largest rebound effects. For example, Lovins and Lovins (1997) used case studies to argue that better visual, acoustic and thermal comfort in well-designed, energy-efficient buildings can improve labour productivity by as much as 16%. Since labour costs in commercial buildings are typically twenty-five times greater than energy costs, the resulting cost savings can dwarf those from reduced energy consumption. But if the total cost savings are twenty-five times greater, the indirect rebound effects may be twenty-five times greater as well.
Saunders (2008) shows how the predicted magnitude of rebound effects depends almost entirely on the choice of the relevant production function—whether at the firm, sector or economy-wide level. Several commonly used production functions are found to be effectively useless in investigating the rebound effect, since the relevant results are the same for whatever values are chosen for key parameters. One popular production function (the constant elasticity of substitution, or CES), is found to be able to simulate rebound effects of different magnitudes, but only if a particular assumption is made about how different inputs are combined. Since this form is widely employed within energy-economic models, Saunders’ results raise serious concerns about the ability of such models to accurately simulate rebound effects. An alternative and more flexible functional form (the ‘Translog’) that is widely used in empirical studies is also found to lead to backfire once standard restrictions are imposed on the parameter values to ensure that the behaviour of the function is consistent with economic theory (Saunders, 2008).²

There is a substantial empirical literature estimating the parameters of different types of production function at different levels of aggregation and obtaining a good fit with observed data. Hence, if such functions are considered to provide a reasonable representation of real-world economic behaviour, Saunders’ work suggests that ‘pure’ energy-efficiency improvements are likely to lead to backfire. Alternatively, if rebound effects are considered to vary widely in magnitude between different sectors, Saunders’ work suggests that standard and widely used economic methodologies cannot be used to simulate them.

The above conclusions apply to pure energy-efficiency improvements. But Saunders (2005) also uses numerical simulations to demonstrate the potential for much larger rebound effects when improvements in energy efficiency are combined with improvements in the productivity of other inputs.³ Again, if the validity of the theoretical assumptions is accepted, these results suggest that backfire may be a more common outcome than is conventionally assumed.

Saunders approach is entirely theoretical and therefore severely limited by the assumptions implicit in the relevant models. For instance, technology always comes free, there are only constant returns to scale in production, markets are fully competitive, there is always full employment, qualitative differences in capital and energy are ignored and so on. Indeed, a considerable literature challenges the idea that an ‘aggregate’ production function for the economy as a whole is meaningful concept (Fisher, 1993; Temple, 2006)—although this may not necessarily invalidate the use of such functions for representing the behaviour of individual sectors. A particular weakness is the assumption that technical change is costless and autonomous, without explicit representation of the processes that affect its rate and direction. This characteristic limits the capacity of such models to address many policy-relevant questions. More recent developments in so-called ‘endogenous growth theory’ have overcome this weakness to some extent, but to date no authors have used such models to explore the rebound effect. However, since what are at issue are the consequences of energy-efficiency improvements, the source of those improvements is arguably a secondary concern.

Overall, Saunders work suggests that significant rebound effects can exist in theory, backfire is quite likely and this result is robust to different model assumptions. Since these results derive from a contested theoretical framework, they are suggestive rather than definitive. But they deserve to be taken seriously.

7 Restrictions normally have to be imposed upon the parameter values in a Translog cost function to ensure that its behaviour is consistent with basic economic theory. In particular, the cost function must be concave, implying that the marginal product of each input declines with increasing use of that input. In many applications, such as CGE modelling, these conditions need to be satisfied for all input combinations, but empirically estimated cost functions sometimes violate these conditions (Diewert and Wales, 1987). Saunders (2008) finds that imposing a global concavity restriction means that the Translog production function always leads to backfire. However, Ryan and Wales (2000) show that if concavity is imposed locally at a suitably chosen reference point, the restriction may be satisfied at most all of the data points in the sample. Under these circumstances, the Translog may be able to represent different types of rebound effect for particular data sets—but only if it can be empirically verified that concavity is honoured across the domain of measurement.

Ecological economists have not directly investigated the rebound effect, but their work arguably provides suggestive support for Jevons’ Paradox in much the same way as Schurr’s research on the historical determinants of US productivity growth. Four examples of this work are briefly described below.

First, analysis by Kaufmann (1992, 2004) and others suggests that historical reductions in energy/GDP ratios owe much more to structural change and shifts towards ‘high-quality’ fuels than to technological improvements in energy efficiency (Box 5). By neglecting changes in energy quality, conventional analysts may have come to incorrect conclusions regarding the rate and direction of technical change and its contribution to reduced energy consumption. Kaufmann (1992) suggests that, not only does the energy/GDP ratio reflect the influence of factors other than energy-saving technical change, but these other factors may be sufficient to explain the observed trends. Hence, the observed improvements in the thermodynamic efficiency of individual devices at the micro level do not appear to have significantly contributed to the observed reduction in energy intensity at the macro-level. As with the work of Jorgenson and others, this suggests that the conventional assumptions of energy-economic models may be flawed.

Second, both neoclassical and ecological economists have used modern econometric techniques to test the direction of causality.
Kaufmann (1992) sought to quantify the factors that contributed to changes in the ratio of primary energy consumption (in kWh thermal) to real GDP in France, Germany, Japan and the UK during the period 1950–1990. The explanatory variables were the percentage share of different energy carriers in primary energy consumption; the fraction of GDP spent directly on energy by households; the proportion of the product mix that originated in energy intensive manufacturing sectors; and primary energy prices.

Despite the simplicity of this formulation, it was found to account for most of the variation in energy intensity for the four countries studied throughout the post-war period. Kaufmann argued that improvements in energy quality led to lower energy intensities by allowing more useful work to be obtained from each heat unit of energy input. The shift from coal to oil contributed greatly to declining energy/GDP ratios prior to 1973, while the rising contribution of primary electricity (hydro and nuclear) provided a significant contribution after 1973.

Since the energy intensity of household energy purchases is an order of magnitude greater than the energy intensity of other goods and services, falls in the former as a fraction of total expenditure should translate into falls in the energy/GDP ratio—and vice versa. The fraction of GDP spent directly on energy by households increased prior to 1973 and decreased thereafter and these trends were also found to be highly significant in explaining trends in the aggregate ratio.

In addition, changes in energy prices encouraged substitution between inputs, including the substitution of capital for energy, while shifts towards less energy-intensive manufacturing sectors and towards the service sector reduced energy/GDP ratios. These mechanisms were found to be less important than those above, but when all four factors were taken into account, they were found to provide a more or less sufficient explanation for the observed trends in aggregate energy intensity.

By implication, Kaufmann’s results suggest little role for energy-saving technical change-defined as advances in technology that allow the same type and quantity of output to be produced with less energy inputs. Kaufmann tested this implication in three different ways (Note: Namely: (a) seeking evidence for serial correlation and heteroscedasticity in the error term, which could be evidence of missing variable bias; (b) including a time trend to represent energy-saving technical change; and (c) using dummy variables to test for changes in the intercept or slope of individual regression coefficients during different time periods—such as may follow an increase in energy prices if this induces energy saving technical change.), but in each case failed to find statistically significant evidence for energy saving technical change. Kaufmann comments that “…Technical changes has reduced the amount of energy (as measured in heat units) used to produce a unit of output. But characterising that technical change as ‘energy-saving’ is misleading. Over the last 40 years, technical change has reduced the amount of energy use to produce a unit of output by developing new techniques for using oil, natural gas and primary electricity in place of coal.”

Kaufmann also interprets the results as illustrating the limited scope, at the level of the macro-economy, for substituting capital and labour for energy. Estimated annually, the own price elasticity of energy demand varies between –0.05 and –0.38, which is generally smaller than the elasticities estimated at the level of individual sectors. This arguably suggests that the indirect energy consumption associated with labour and capital inputs constitutes a significant portion of the energy saved directly through energy efficiency improvements in each of those sectors.

Table 2
Trends in second-law conversion efficiencies of primary conversion processes in the US (average percentage efficiency in specified year).

<table>
<thead>
<tr>
<th>Year</th>
<th>Electricity generation and distribution</th>
<th>Transportation</th>
<th>High-temperature process heat (steel)</th>
<th>Medium-temperature process heat (steam)</th>
<th>Low-temperature space heat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1900</td>
<td>3.8</td>
<td>7</td>
<td>5</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>1970</td>
<td>32.5</td>
<td>20</td>
<td>14</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1990</td>
<td>33.3</td>
<td>25</td>
<td>20</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Source: Ayres et al. (2003).

between energy consumption and GDP (Chontanawat et al., 2006; Lee, 2006; Lee and Chang, 2005; Stern, 1993; Yoo, 2005; Zachariadis, 2006). The assumption is that if GDP growth is the cause of increased energy consumption then a change in the GDP growth rate should be followed by a change in energy consumption and vice versa. It is argued that if causality runs from GDP to energy consumption then energy consumption may be reduced without adverse effects on economic growth, while if causality runs the other way round a reduction in energy use may negatively affect economic growth. While the results of such studies are frequently contradictory, most of them neglect changes in energy quality. When energy quality is taken into account, the causality appears to run from energy consumption to GDP—as ecological economists suggest (Stern, 1993, 2000).

Third, historical experience provides very little support for the claim that increases in income will lead to declining energy consumption (Richmond and Kaufmann, 2006b; Stern, 2004a; Stern and Cleveland, 2004). While the income elasticity of aggregate energy consumption may be both declining and less than one in OECD countries, there is no evidence that it is negative (or is soon to become negative). Again, neglect of changes in fuel mix and energy prices may have led earlier studies to draw misleading conclusions regarding the extent to which energy consumption has been decoupled from GDP (Kaufmann, 2004).

Finally, ecological economists have developed a number of alternatives to the conventional models of economic growth (Ayres, 1998; Ayres and Warr, 2002a; Ayres and van den Bergh, 2005; Ayres and Warr, 2005a, 2006; Beaudreau, 1995a, b, 1998; Beaudreau, 2005; Kummel, 1982, 1989; Kummel et al., 2002, 2000, 1985). A key feature of these models is a departure from the traditional assumption that the productivity of each input is proportional to the share of that input in the value of output. Instead, the productivity of each input is estimated directly from a production function. These models are found to reproduce historical trends in economic growth extremely well, without attributing any role to technical change. This is in contrast to conventional theories of economic growth, which attribute much
of the increase in output to technical change. The marginal productivity of energy inputs is found to be around ten times larger than their cost share, implying that improvements in energy productivity could have a dramatic effect on economic growth and therefore on economy-wide energy consumption—in other words, the rebound effect could be very large.

Of particular interest is the work by Ayres and Warr (2005a), who combine historical data on the ‘exergy’ content of fuel inputs and second-law thermodynamic conversion efficiencies to develop a unique time series of the exergy output of conversion devices (termed useful work) in the US economy over the past century. They show that useful work inputs to the US economy have grown by a factor of 18 over the past 100 years, implying that the useful work obtained from fuel resources has grown much faster than the consumption of fuels themselves, owing to substantial improvements in thermodynamic conversion efficiencies. By including useful work in their production function, rather than primary energy, Ayres and Warr (2005a,b) obtain an extremely good fit to US GDP trends over the past century, thereby eliminating the need for a multiplier for technical change. The implication is that improvements in thermodynamic conversion efficiency provide a quantifiable surrogate for all forms of technical change that contribute to economic growth. Far from being a minor contributor to economic growth, improvements in thermodynamic efficiency become the dominant driver—obviating the need for alternative measures of technological change.

The implication of this work is that energy is more productive than is suggested by its small share of total costs. This is precisely the argument that Schurr made and which appears to underlie some of Brookes’ arguments in favour of backfire. However, the empirical evidence in support of this perspective remains patchy. For example, the results of econometric investigations of causality relationships between energy and GDP remain ambiguous and the policy implications that are drawn are frequently oversimplified (Zachariadis, 2006). Also, the statistical form of causality that is being measured here (so-called ‘Granger causality’) is not the same as causality as conventionally understood and conventional notions of causality may be problematic for systems as complex as modern economies (as Fig. 3 indicates). In a similar manner, the different variants of ‘ecological growth models’ rely upon an unusual and oddly behaved production function, provide results that are difficult to reconcile with each other and appear vulnerable to bias from a number of sources that could potentially invalidate the results (Sorrell and Dimitropoulos, 2007). As a result, claims that the marginal productivity of energy is an order of magnitude larger than its cost share, or that improvements in thermodynamic conversion efficiencies can act as a suitable proxy for technical change, must be treated with considerable caution.

Unfortunately, the different assumptions of conventional and ecological perspectives on economic growth seem to have prevented an objective comparison of their methods and conclusions. Convincing evidence of the disproportionate contribution of energy to productivity improvements and economic growth, therefore, remains elusive. Moreover, even if this were to be accepted, the link from this evidence to Jevons’ Paradox remains ambiguous and indirect.

The neoclassical assumption appears to be that capital, labour and energy inputs have independent and additive effects on economic output, with any residual increase being attributed to exogenous technical change. Endogenous growth theory has modified these assumptions, but still attributes a relatively minor role to energy. In contrast, the ecological assumption appears to be that capital, labour and energy are interdependent inputs that have synergistic and multiplicative effects on economic output, and that the increased availability of low-cost, high-quality energy sources has provided a necessary condition for most historical improvements in economic productivity. A bridge between the two could potentially be provided by Toman and Jemelkova’s (2003) observation that increased inputs of useful work (or energy services) may enhance the productivity of capital and labour:

……when the supply of energy services is increased, there is not just more energy to be used by each skilled worker or machine; the productivity with which every unit of energy is used also rises. If all inputs to final production are increased in some proportion, final output would grow in greater proportion because of the effect on non-energy inputs. (Toman and Jemelkova, 2003).

The account of Schurr (1983, 1984, 1985) for the impact of electricity (and especially electric motors) on the organisation and productivity of US manufacturing provides an example of this process. But the extent to which such patterns have applied in other sectors and time periods needs to be determined empirically. If such a situation is the norm, the increased availability of high-quality energy may be a primary driver of economic activity. But if the increased availability of high-quality energy inputs has a disproportionate impact on productivity and economic growth, then improvements in thermodynamic conversion efficiency may do the same, because both increase the useful work available from conversion devices. Under these conditions, the rebound effect could be large and potentially greater than unity.

8. Implications

Jevons’ Paradox implies that all economically justified energy-efficiency improvements will increase energy consumption above where it would be without those improvements. Since this is a counterintuitive claim for many people, it requires strong supporting evidence if it is to gain widespread acceptance. The main conclusion from this paper is that such evidence does not yet exist. The theoretical and empirical evidence cited in favour of the Paradox contains a number of weaknesses and inconsistencies and most is only indirectly relevant to the rebound effect. Nevertheless, the arguments and evidence deserve more serious attention than they have received to date. Much of the evidence points to economy-wide rebound effects being larger than is conventionally assumed and to energy playing a more important role in driving productivity improvements and economic growth than is conventionally assumed.

The possibility of large economy-wide rebound effects has been dismissed by a number of leading energy analysts (Howarth, 1997; Laitner, 2000; Lovins, 1988, 1998; Schipper and Grubb,
2000). But it becomes more plausible if it is accepted that energy-efficiency improvements are frequently associated with improvements in the productivity of other inputs. If this is the case, then rebound effects need not necessarily be small just because the share of energy in total costs is small. This possibility is suggested by both Brooke’s work and that of the ecological economists and is also demonstrated in a neo-classical framework by Saunders (1992a, 2005). Future research should, therefore, investigate the extent to which improvements in energy efficiency (however defined and measured) are associated with broader improvements in economic productivity, and the circumstances under which economy-wide rebound effects are more or less likely to be large (Fig. 4).

Rebound effects may be particularly large for the energy-efficiency improvements associated with so-called ‘general-purpose technologies’, such as steam engines and computers. General-purpose technologies (GPTs) have a wide scope for improvement and elaboration, are applicable across a broad range of uses, have potential for use in a wide variety of products and processes and have strong complementarities with existing or potential new technologies (Lipsey et al., 2005). Steam engines provide a paradigmatic illustration of a GPT in the nineteenth century, while electric motors provide a comparable illustration for the early 20th century. The former was used by Jevons to support the case for backfire, while the latter formed a key component of Schurr’s work which was subsequently cited by Brookes as suggestive evidence for backfire.

The key to unpacking Jevons’ Paradox may therefore be to distinguish the energy-efficiency improvements associated with GPTs from other forms of energy-efficiency improvement. Jevons’ Paradox seems more likely to hold for the former, particularly when these are used by producers and when the energy-efficiency improvements occur at an early stage of development and diffusion of the technology. The opportunities offered by these technologies have such long term and significant effects on innovation, productivity and economic growth that economy-wide energy consumption is increased. In contrast, Jevons’ Paradox seems less likely to hold for dedicated energy-efficiency technologies such as improved thermal insulation, particularly when these are used by consumers or when they play a subsidiary role in economic production. These technologies have smaller effects on productivity and economic growth, with the result that economy-wide energy consumption may be reduced.

The implication is that climate policy should focus on encouraging dedicated energy-efficiency technologies, rather than improving the energy efficiency of GPTs. However, these categories are poorly defined and the boundaries between them are blurred. Moreover, even if GPTs can meaningfully be distinguished from other forms of technology, continued economic growth is likely to depend upon the diffusion of new types of GPT that may increase aggregate energy consumption.

9. Summary

The case for Jevons’ Paradox is not based upon empirical estimates of rebound effects. Instead, it relies largely upon theoretical arguments, backed up by empirical evidence that is both suggestive and indirect. Disputes over the Paradox, therefore, hinge in part on competing theoretical assumptions. While historical experience demonstrates that substantial improvements in energy efficiency have occurred alongside increases in economic output, total factor productivity and overall energy consumption, this does not provide sufficient evidence for Jevons’ Paradox since the causal links between these trends remains unclear.

This paper has reviewed the strengths and weaknesses of the arguments and evidence in favour of backfire presented by Len Brookes and Harry Saunders. While neither author provides a totally convincing case in favour of Jevons’ Paradox, their work poses an important challenge to conventional wisdom. Of particular interest is the apparent similarity between some of Brookes’ arguments and the heterodox claim that the increased availability and quality of energy inputs is the primary driver of economic activity. But energy is only economically productive because it provides useful work. Hence, if increases in energy inputs contribute disproportionately to economic growth, then improvements in thermodynamic efficiency could do the same, since both provide more useful work. The dispute over Jevons’ Paradox may therefore be linked to a broader question of the contribution of energy to economic growth.

The debate over Jevons’ Paradox would benefit from further distinctions between different types of energy-efficiency improvement. In particular, Jevons’ Paradox seems more likely to hold for...
energy-efficiency improvements associated with the early stage of diffusion of ‘general-purpose technologies’, such as electric motors in the early twentieth century. It may be less likely to hold for the later stages of diffusion of these technologies, or for ‘dedicated’ energy-efficiency technologies such as improved thermal insulation. However, these categories are poorly defined, the boundaries between them are blurred and GPTs account for a significant portion of total energy consumption.

Overall, while it is seems unlikely that all energy-efficiency improvements will lead to backfire, we still have much to learn about the factors that make backfire more or less likely to occur. This review suggests several possible avenues for research, which may supplement attempts to quantify rebound effects. First, econometric and decomposition techniques could be used to better understand the source of changes in aggregate energy efficiency (e.g. the relative contribution of structural change, technical change, input substitution, changes in fuel mix and other factors) (Sue Wing, 2008). Second, these techniques could also be used to investigate the extent to which different types of energy efficiency improvement are associated with improvements in the productivity of other inputs and with improvements in total factor productivity. Third, the implications of Saunder’s work on neoclassical production theory could be further assessed, especially since the relevant functions underpin most standard energy-economic models. Finally, the ecological models of economic growth need more critical scrutiny, both to assess their statistical robustness and to explore whether and how they can be reconciled with more conventional approaches. While communication is inhibited in part by competing ‘world-views’, there should be scope for mutual learning and improved testing.

A prerequisite for all the above is a recognition that rebound effects matter and need to be taken seriously. Something is surely amiss when such in-depth and comprehensive studies as the Stern (2007) review overlook this topic altogether. While rebound effects are difficult to study, they are not necessarily any more difficult than well-researched issues such as price-induced technical change. Their continued neglect may result as much from their uncomfortable implications as from a lack of methodological tools. Too much is at stake for this to continue.

Acknowledgements

This paper is based upon a comprehensive review of the evidence for rebound effects, conducted by the UK Energy Research Centre (Sorrell, 2007). Earlier versions of this paper are contained in Hering and Sorrell (2008) and Sorrell (2007), while the research upon which it is based is reported in detail in Sorrell and Dimitropoulos (2007). The latter reports are available to download from the UKERC website. The financial support of the UK Research Councils is gratefully acknowledged.

Considerable thanks are due to John Dimitropoulos (Mott MacDonald) for his contribution to this work and also to Harry Saunders (Decision Processes Inc.) for his extremely helpful comments and encouragement. The author is also very grateful for the advice and comments received from Manuel Frondel, Karsten Neuhoff, Jake Chapman, Nick Eyre, Serban Scriciu, Blake Alcott, Len Brookes, John Feather, Gordon Mackerron, Horace Herring, Lester Hunt and Paolo Aognolucci, Jim Skea, Rob Gross and Phil Heptonstall, as well from two anonymous referees. The usual disclaimers apply.

References


Beaudreau, B.C., 2005. Engineering and economic growth. Structural Change and Economic Dynamics Contains the Special Issue Approaches to Production Theory 16 (2), 211–220.


Communication

Global oil depletion: A review of the evidence

Steve Sorrell a,⁎, Jamie Speirs b, Roger Bentley c, Adam Brandt d, Richard Miller e

a Sussex Energy Group, SPRU (Science and Technology Policy Research), Freeman Centre, University of Sussex, Falmer, Brighton BN1 9QE, UK
b Imperial College Centre for Energy Policy and Technology, UK
c Department of Cybernetics, University of Reading, UK
d Department of Energy Resources Engineering, Stanford University, USA
e Independent Consultant, UK

A R T I C L E   I N F O

Article history:
Received 9 April 2010
Accepted 23 April 2010
Available online 20 May 2010

Keywords:
Depletion
Peak oil
Oil supply forecasting

A B S T R A C T

Within the polarised and contentious debate over future oil supply a growing number of commentators are forecasting a near term peak and subsequent decline in production. But although liquid fuels form the foundation of modern industrial economies, the growing debate on 'peak oil' has relatively little influence on energy and climate policy. With this in mind, the UK Energy Research Centre (UKERC) has conducted an independent, thorough and systematic review of the evidence, with the aim of establishing the current state of knowledge, identifying key uncertainties and improving consensus. The study focuses upon the physical depletion of conventional oil in the period to 2030 and includes an in-depth literature review, analysis of industry databases and a detailed comparison of global supply forecasts. This Communication summarises the main findings of the UKERC study. A key conclusion is that a peak of conventional oil production before 2030 appears likely and there is a significant risk of a peak before 2020.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

Conventional oil (namely crude oil, condensate and natural gas liquids) currently accounts for over 97% of global liquid fuels production and is still expected to account for around 90% in 2030 (IEA, 2008). But many commentators are forecasting a near-term peak and subsequent terminal decline in the production of conventional oil, with alternative sources being unable to 'fill the gap' on the timescale required (Aleklett et al., 2009; Campbell and Heapes, 2008). In contrast, others argue that liquid fuels production will be sufficient to meet global demand well into the 21st century, as rising prices stimulate new discoveries, enhanced recovery and the development of non-conventional resources such as oil sands (Adelman, 2003; Mills, 2008; Odell, 2004). While the global economic recession has changed the short-term outlook, many are forecasting renewed supply constraints around 2013 (Blanch et al., 2010; ITPOES, 2010).

Given the polarised, contentious and confused nature of the peak oil debate, the UK Energy Research Centre (UKERC) chose to undertake an independent, thorough and systematic review of the evidence, with the aim of establishing the current state of knowledge, identifying key uncertainties and improving consensus. The project focused upon the physical depletion of conventional oil in the period to 2030 and included an in-depth review of more than 500 studies, the analysis of industry databases and a detailed comparison of 14 global supply forecasts (Sorrell et al., 2009). The following summarises the main findings.

2. Context

Conventional oil is defined here to include crude oil, condensate and natural gas liquids (NGLs) and to exclude liquid fuels derived from oil sands, oil shale, natural gas and biomass. The future supply of conventional oil is constrained by three physical features of the resource:

• Production from individual fields normally rises to a peak or plateau, after which it declines as a result of falling pressure and/or the breakthrough of water. Decline normally begins when less than half of the recoverable resources of a field have been produced (Höök et al., 2009b; IEA, 2008).

1 Crude oil comprises a mixture of hydrocarbons that exist in liquid phase in underground reservoirs and which remain liquid at surface temperature and pressure. In contrast, condensate exists as gas in underground reservoirs but condenses to liquids at the surface. Natural gas liquids (NGLs) include condensate together with lighter hydrocarbons such as butane and propane which can be liquefied with the application of moderate pressure. The condensate produced from associated gas at oil wells is typically combined with crude oil in production data, while NGL production is recorded separately and is largely a by-product of natural gas production. On an energy equivalent basis, the current ratio of NGL to gas production is ~15% (BP 2008; EIA, 2007). Gas resources are less developed than oil resources, so physical depletion is less of a constraint on future NGL production. But only a portion of NGLs can be blended into transport fuels.
Most oil in a region tends to be located in a small number of large fields, with the balance being located in a much larger number of small fields. These large fields tend to be discovered relatively early, in part because they occupy a larger area. Subsequent discoveries tend to be progressively smaller and often require more effort to locate.

At some point, the additional production from small fields that are discovered relatively late will become insufficient to compensate for the decline in production from large fields that were discovered relatively early, leading to a regional peak in production (Fig. 1). This pattern has been observed for over one hundred oil-producing regions around the world and will ultimately be repeated at the global level. While the timing and shape of this peak is subject to multiple technical, economic and political influences, the range of possibilities becomes more constrained as the resource is progressively depleted.

Global cumulative production of conventional oil stood at \( \approx 1128 \text{ billion barrels (Gb)} \) in 2007, with annual production of 29.5 Gb. Since 1995, global production has grown at an average of 1.5%/year, with 60% of cumulative production occurring since 1980 (Fig. 2). Most of the world’s conventional oil was discovered between 1946 and 1980 (Fig. 3) and since that time annual production has exceeded annual discoveries (Fig. 4). However, annual production is less than annual reserve additions since the latter includes upward revisions to the reserve estimates for known fields (‘reserve growth’). Using industry data on proved and probable (2P) reserves, we estimate that between 2000 and 2007 an average of 48 Gb was added to global reserves each year, split between 15 Gb/year of new discoveries and 33 Gb/year of reserve growth. However, these estimates are uncertain and contested and many expect the rate of reserve additions to decline.

3. Data quality and interpretation

Publicly available data sources are poorly suited to studying oil depletion and their limitations are insufficiently appreciated (Bentley et al., 2007). The databases available from commercial sources are better in this regard, but are also expensive, confidential and not necessarily reliable for all regions. While terms such as ‘proved’ (1P) and ‘proved and probable’ (2P) reserves are widely used, these are defined and interpreted in different ways with only limited progress towards standardisation. Only a subset of global reserves is subject to formal reporting requirements and this is largely confined to the reporting of highly conservative 1P data for aggregate regions. In the absence of audited estimates for individual fields, analysts must rely upon assumptions whose level of confidence is inversely proportional to their importance – being lowest for those countries that hold the majority of the world’s reserves.

The common practice of adding 1P estimates to form a regional or global total is statistically incorrect and likely to significantly underestimate actual 1P reserves. Aggregation of 2P estimates should introduce less error but this may be either positive, negative or zero depending upon the interpretation of the estimates and the shape of the underlying probability distributions. As a result of inappropriate aggregation, global 1P reserves could be larger than the 1240 Gb reported by BP (2008) – potentially offsetting the overestimation of OPEC reserves that is claimed by some authors (Campbell and Heapes, 2008). At the same time, the industry estimates of global 2P reserves are approximately the same, suggesting either that the former have been overestimated or the latter underestimated. Since the discrepancies between these data sources vary both in magnitude and sign from one country to another, the global totals should be treated with considerable caution.

4. Key variables

While the observed lognormal size distribution of discovered fields is partly the result of sampling bias, there is insufficient evidence to conclude whether a ‘linear fractal’ or ‘parabolic fractal’ better describes the population size distribution (Laherrère, 2000). But the number of small fields is of secondary importance: while technical improvements and higher prices should make more of these fields viable, their exploitation will be subject to rapidly diminishing returns. Although there are around 70,000 producing oil fields in the world, approximately 25 fields account for one quarter of global production, 100 fields account for half of production and up to 500 fields account for two-thirds of production.
Most of these ‘giant’ fields are relatively old, many are well past their peak of production, most of the rest will begin to decline within the next decade or so and few new giant fields are expected to be found. The remaining reserves at these fields, their future production profile and the potential for enhanced recovery and reserve growth are therefore of critical importance.

Estimates of the recoverable resources of individual fields tend to increase over time as a result of improved geological knowledge, better technology, changes in economic conditions and revisions to initially conservative reserve estimates. This process is currently adding more to global reserves each year than the discovery of new fields and it seems likely to continue to do so in the medium-term. The contribution of different factors to ‘reserve growth’ varies widely between different fields, regions and datasets and is neither solely or necessarily largely the result of conservative reporting. Since 1995, the average growth in 2P reserve estimates seems broadly in line with the optimistic assumptions made by the US Geological Survey (USGS, 2000), but the global average is strongly influenced by reserve growth in those countries with the largest reserves and the poorest quality data. Also, reserve growth appears to be greater for larger, older and onshore fields, so as global production shifts towards newer, smaller and offshore fields the rate of growth may decrease in both percentage and absolute terms.

The oil industry must continually invest to replace the decline in production from existing fields. The production-weighted global average rate of decline from post-peak fields is at least...
6.5%/year, while the corresponding rate of decline from all currently-producing fields is at least 4%/year. This implies that at least 3 mb/d of new capacity must be added each year, simply to maintain production at current levels. Decline rates are on an upward trend as more giant fields enter decline, as production shifts towards smaller, younger and offshore fields (which decline faster) and as changing production methods lead to more rapid post-peak decline (e.g. Höök et al., 2009a, 2009b). As a result, more than two-thirds of current crude oil production capacity may need to be replaced by 2030, simply to prevent production from falling.

There are physical, engineering and economic constraints upon both the rate of depletion of a field or region and the pattern of production over time. For example, the annual production from a region has rarely exceeded 5% of the remaining recoverable resources and most regions have reached their peak well before half of their recoverable resources have been produced. If the global average depletion rate for newly developed resources is as high as the maximum depletion rate previously seen in any oil-producing region, at least 20 Gb must be added to global reserves each year to compensate for the decline in production from existing fields. With lower depletion rates, accelerating decline rates and growing demand, a significantly higher rate of reserve additions will be required.

5. Methods of estimating resource size

The ultimately recoverable resources (URR) of a region depend upon economic and technical factors as much as geology and can only be estimated to a reasonable degree of confidence when exploration is well advanced. There are a variety of methods for estimating URR, with 'geological' techniques being more appropriate for relatively unexplored regions and techniques based upon the extrapolation of production or discovery trends being more appropriate where exploration is advanced. The confidence bounds on most estimates are relatively large and the few studies that compare different techniques show that they can lead to very different results (Ahlbrandt and Klett, 2005).

Simple ‘curve-fitting’ techniques require only aggregate data on regional production or discoveries, but are best applied to geologically homogeneous regions with a relatively unrestricted exploration history. Many applications of these techniques take insufficient account of their limitations, including their inadequate theoretical basis, the sensitivity of the estimates to the choice of functional form, the inability to anticipate future cycles of production or discovery and the implications of neglecting economic and other variables. Curve fitting to discovery data introduces additional complications such as the uncertainty in reserve estimates and the need to adjust these estimates to allow for future reserve growth. In general, these weaknesses are more likely to lead to underestimates of the URR and excessively pessimistic forecasts of oil supply. Different techniques, functional forms, length of time series and numbers of curves can lead to very different results, but the degree of uncertainty declines as exploration matures. Hybrid models that incorporate economic and political variables provide a better fit to historical data, but may not lead to substantially different estimates of the regional URR.

6. Methods of forecasting future supply

Methods of supply forecasting vary widely in terms of their theoretical basis, inclusion of different variables, level of aggregation and complexity. Each approach has its strengths and weaknesses and none should be favoured in all circumstances. Curve-fitting models are straightforward and widely used, but lack an adequate theoretical basis; are sensitive to the choice of functional form, neglect key variables and can perform poorly as a result. Econometric models provide a better match to historical data, but this may not translate to more accurate forecasts of future production. Hybrids of curve-fitting and econometrics offer promise, but can also have the disadvantages of both. Systems dynamic models can reproduce the physical and economic mechanisms that govern oil production, but can also be over-complicated and unstable and frequently lack both empirical validation and sufficient data for parameterisation. Finally, bottom-up models using field or project data may provide the most reliable basis for near-term forecasts, but existing models are hampered by their reliance on proprietary datasets, lack of transparency, uncertainty over key variables and the need to make multiple assumptions.

The timing of the global peak can be estimated to within decadal accuracy assuming a particular value for global URR and no significant disruptions to the global oil market. But given the potential for political, economic, or technological disruptions, no model can provide estimates of great precision. Increasing model complexity does little to address this problem and is subject to rapidly diminishing returns.

7. The global ultimately recoverable resource

Estimates of the global URR for conventional oil vary widely in their definitions, methods, assumptions and results. Although such estimates have been trending upwards for the last 50 years, the mean estimate of 3345 Gb from the USGS (2000) represents a substantial departure from the historical trend. Contemporary estimates now fall within the range 2000–4300 Gb, compared to cumulative production through to 2007 of 1128 Gb. But despite their apparent optimism, assertions that the USGS estimates are ‘discredited’ are premature. Global reserve growth appears to be matching the USGS assumptions and although the rate of new discoveries is much lower than implied by the USGS mean estimate, the size of these discoveries may have been underestimated and exploration remains restricted in key areas.

The timing of the global peak for conventional oil production is relatively insensitive to assumptions about the size of the resource (e.g. see Fig. 5). For a wide range of assumptions about the global URR and the shape of the future production cycle, a peak in production can be estimated to occur before 2031 (Kauffman and Shiers, 2008). In most models, increasing the global URR by a billion barrels delays the peak by only a few days. Delaying the peak beyond 2030 requires optimistic assumptions about the size of the recoverable resource and the rate at which it is developed, combined with a slow rate of demand growth prior to the peak and/or a relatively steep decline in production following the peak. These considerations constrain the range of plausible supply forecasts.

Much of the remaining recoverable resource is located in smaller fields in less accessible locations. If (as seems likely) these resources can only be produced relatively slowly at high cost, supply constraints may inhibit demand growth at a relatively early stage. Demand growth may also be constrained if the national oil companies that control much of these resources lack the incentive or ability to invest.

---

2 This was updated to 3577 Gb, by the IEA (2008) but the inclusion of ‘conventional oil produced by unconventional means’ increases this to 4276 Gb. However, the latter estimate relies upon a large contribution from enhanced oil recovery that the IEA anticipate will take decades to be realised.
production data, constrained by different assumptions about the global URR.

These include the following:

- optimistic and at worst implausible. These include the following:
- after 2030 rest upon several assumptions that are at best forecasts that delay the peak of conventional oil production until outcome than a sharp peak. Nevertheless, we consider that and demand make a multi-year ‘bumpy plateau’ a more likely peak production unwarranted. Also, interactions between supply uncertainties multiply making precise forecasts of the timing of medium to long-term forecasting, the number and scale of consequent risk of supply constraints when demand recovers. For and delay of these projects as a result of the recession and the planned. The primary issue for the short term is the cancellation supply have long lead times and most are already committed or is relatively inflexible, because the projects which will raise priority should be given to constraining these parameters to a greater degree than at present. The short term future of oil production capacity, to about 2016, is relatively inflexible, because the projects which will raise supply have long lead times and most are already committed or planned. The primary issue for the short term is the cancellation and delay of these projects as a result of the recession and the consequent risk of supply constraints when demand recovers. For medium to long-term forecasting, the number and scale of uncertainties multiply making precise forecasts of the timing of peak production unwarranted. Also, interactions between supply and demand make a multi-year ‘bumpy plateau’ a more likely outcome than a sharp peak. Nevertheless, we consider that forecasts that delay the peak of conventional oil production until after 2030 rest upon several assumptions that are at best optimistic and at worst implausible. These include the following:
  - lower rates of demand growth than are currently assumed (e.g. 1%/year);
  - a global URR that is comparable to or greater than the mean estimate of the USGS (∼3600 Gb);
  - a rapid decline in production following the peak (e.g. 3%/year or more);
  - cumulative production at the date of peak that exceeds 50% of the global URR (i.e. much greater than previously observed in the majority of post-peak regions);
  - cumulative production at the date of peak that exceeds 60% of cumulative 2P discoveries (i.e. much greater than previously observed in the majority of post-peak regions);
  - an annual rate of new discoveries over the period the 2030 that equals or exceeds that achieved over the last decade (i.e. reversing the trend of the last 40 years, despite the declining size of newly discovered fields);
  - an annual rate of reserve growth over the period to 2030 that equals or exceeds that achieved over the last decade (despite the growing share of newer, smaller and offshore fields that have less potential for reserve growth);
  - depletion of these resources at an average rate that is much higher than the maximum rate previously achieved in any oil-producing region (Aleklett et al., 2009) and favourable ‘above-ground’ conditions, including appropriate incentives for investment, sufficient access to prospective areas, political stability and so on in all the major oil-producing regions.

Such forecasts need to either demonstrate how these assumptions can be met or why they do not apply. On the basis of current evidence we suggest that a peak of conventional oil production before 2030 appears likely and there is a significant risk of a peak before 2020.

A peak in conventional oil supply will only be associated with a peak in liquid fuels supply if ‘non-conventional’ sources are unable to substitute in a sufficiently timely fashion. However, numerous technical, economic and environmental constraints make a rapid expansion of non-conventional production extremely challenging. For example, Söderbergh et al. (2007) estimate that a ‘crash programme’ to develop the Canadian oil sands could deliver only 5 mb/d by 2030 while CERA (2009) estimate a maximum of 6 mb/d. This is less than 6% of the IEA (2008) projection of global liquid fuels demand in 2030, with other non-conventional sources being projected to deliver much smaller volumes (Bentley et al., 2009; IEA, 2008). There should be greater scope for mitigation on the demand side, but this also requires large-scale investment in multiple options with long lead times (Hirsch et al., 2005; McCollum and Yang, 2009). Hence, the risk of a peak in conventional oil production deserves urgent and serious consideration.

8. The global supply outlook

Comparison of global oil supply forecasts is hampered by the lack of transparency of many of the models, the inconsistency in the definition and coverage of liquids and the range of methods and assumptions used. There is considerable scope for improving consensus by revealing and comparing key assumptions and systematically exploring the sensitivity of forecasts to those assumptions. Models need to better integrate supply and demand and explore a wider range of socioeconomic scenarios.

Although the range of modelling approaches, assumptions and results is very wide, global forecasts of conventional oil supply differ largely in their explicit or implicit assumptions for the global URR and/or post-peak aggregate production decline rate. It seems likely that most models would give similar results if they used or implied similar values for these parameters. Hence, priority should be given to constraining these parameters to a greater degree than at present.

The short term future of oil production capacity, to about 2016, is relatively inflexible, because the projects which will raise supply have long lead times and most are already committed or planned. The primary issue for the short term is the cancellation and delay of these projects as a result of the recession and the consequent risk of supply constraints when demand recovers. For medium to long-term forecasting, the number and scale of uncertainties multiply making precise forecasts of the timing of peak production unwarranted. Also, interactions between supply and demand make a multi-year ‘bumpy plateau’ a more likely outcome than a sharp peak. Nevertheless, we consider that forecasts that delay the peak of conventional oil production until after 2030 rest upon several assumptions that are at best optimistic and at worst implausible. These include the following:

- lower rates of demand growth than are currently assumed (e.g. 1%/year);
- a global URR that is comparable to or greater than the mean estimate of the USGS (∼3600 Gb);
- a rapid decline in production following the peak (e.g. 3%/year or more);
- cumulative production at the date of peak that exceeds 50% of the global URR (i.e. much greater than previously observed in the majority of post-peak regions);
- cumulative production at the date of peak that exceeds 60% of cumulative 2P discoveries (i.e. much greater than previously observed in the majority of post-peak regions);
- an annual rate of new discoveries over the period the 2030 that equals or exceeds that achieved over the last decade (i.e. reversing the trend of the last 40 years, despite the declining size of newly discovered fields);
- an annual rate of reserve growth over the period to 2030 that equals or exceeds that achieved over the last decade (despite the growing share of newer, smaller and offshore fields that have less potential for reserve growth);
- depletion of these resources at an average rate that is much higher than the maximum rate previously achieved in any oil-producing region (Aleklett et al., 2009) and favourable ‘above-ground’ conditions, including appropriate incentives for investment, sufficient access to prospective areas, political stability and so on in all the major oil-producing regions.

Such forecasts need to either demonstrate how these assumptions can be met or why they do not apply. On the basis of current evidence we suggest that a peak of conventional oil production before 2030 appears likely and there is a significant risk of a peak before 2020.

A peak in conventional oil supply will only be associated with a peak in liquid fuels supply if ‘non-conventional’ sources are unable to substitute in a sufficiently timely fashion. However, numerous technical, economic and environmental constraints make a rapid expansion of non-conventional production extremely challenging. For example, Söderbergh et al. (2007) estimate that a ‘crash programme’ to develop the Canadian oil sands could deliver only 5 mb/d by 2030 while CERA (2009) estimate a maximum of 6 mb/d. This is less than 6% of the IEA (2008) projection of global liquid fuels demand in 2030, with other non-conventional sources being projected to deliver much smaller volumes (Bentley et al., 2009; IEA, 2008). There should be greater scope for mitigation on the demand side, but this also requires large-scale investment in multiple options with long lead times (Hirsch et al., 2005; McCollum and Yang, 2009). Hence, the risk of a peak in conventional oil production deserves urgent and serious consideration.

References


footnote text: Also, volumetric measures of production are misleading since the net energy yield from non-conventional sources is substantially lower than that from conventional oil.
Oil futures: A comparison of global supply forecasts

Steve Sorrell a,*, Richard Miller b, Roger Bentley c, Jamie Speirs d

a Sussex Energy Group, SPRU (Science and Technology Policy Research), Freeman Centre, University of Sussex, Falmer, Brighton BN1 9QE, UK
b Independent Consultant, UK
c Department of Cybernetics, University of Reading, UK
d Imperial College Centre for Energy Policy and Technology, UK

ABSTRACT
This paper compares and evaluates fourteen contemporary forecasts of global supply of conventional oil and provides some observations on their relative plausibility. Despite the wide range of modelling approaches used and multiplicity of assumptions made, it is shown that forecasts can be usefully compared along two dimensions, namely: shape of future production profile and assumed or implied ultimately recoverable resource of conventional oil. Other differences between forecasts are either secondary or are components of these two parameters. The paper shows how large differences in the assumed size of the resource make relatively little difference to the timing of a global peak in conventional oil production. It also examines the impact of rates of discovery, reserves growth and depletion on the forecast date of peak and shows how forecasts that delay this peak until beyond 2030 rest on assumptions that are at best optimistic and at worst implausible.

Keywords: Depletion, Peak oil, Ultimately recoverable resource

1. Introduction
This paper compares and evaluates some contemporary forecasts of global supply of conventional oil and provides some observations on their relative plausibility. Despite the wide range of modelling approaches used and multiplicity of assumptions made, forecasts can be usefully compared along two dimensions, namely: shape of future production profile and the assumed or implied ultimately recoverable resource (URR) of conventional oil. These features are either explicit or implicit in all the forecasts compared.

The following section provides some necessary background, including physical constraints on conventional oil supply, relevant definitions and recent trends. Section 3 discusses the ultimately recoverable resource (URR) of conventional oil and shows how this constrains the timing of a future peak in global production. Section 4 provides an overview of fourteen contemporary forecasts of global oil supply and summarises their key features and assumptions. Section 5 compares these forecasts in terms of their assumed or implied URR for conventional oil, together with rate of production decline following the peak. Section 6 discusses how the rate of reserve additions from new discovery and reserve growth could impact the forecast timing of peak production. Section 7 summarises the main lessons.

2. Concepts, definitions and trends

Conventional oil is defined here as the combination of crude oil, condensate and natural gas liquids (NGLs: Fig. 1). 1 Conventional oil forms a subset of all oil production, which also includes oil sands and extra heavy oil, and this in turn forms a subset of all-liquids production, which also includes coal to liquids (CTLs), gas to liquids (GTLs) and biofuels. Definitions of these liquids are given by the EIA (2006) and other data sources but these are not always consistent. Global production of conventional oil averaged 80.7 million barrels per day (mb/d) in 2007, 2 split between 70.2 mb/d of crude oil and condensate and 10.5 mb/d of NGLs (Fig. 2). Together, these formed over 97% of the global production of liquid fuels (82.9 mb/d) and the IEA (2008) projects that conventional oil will still provide around 90% of global production in 2030. If crude oil depletes more rapidly than the IEA

1 Crude oil is “… a mixture of hydrocarbons that exist in liquid phase in natural underground reservoirs and which remain liquid at atmospheric temperature and pressure” (EIA, 2006). Condensates are hydrocarbons that exist in gaseous form in the reservoirs but condense to liquids at surface temperatures and pressures. Natural gas liquids (NGLs) are light hydrocarbons from gas reservoirs that are either liquid at atmospheric temperatures and pressures, or can be relatively easily liquefied under moderate pressure. On an energy equivalent basis, the current ratio of NGLs to gas production is approximately 15%. BP, 2008; EIA, 2007). Gas resources are less depleted than crude oil resources, but only a portion of NGLs can be blended into transport fuels.

2 A barrel is equivalent to 159 litres. The weight of a barrel depends on the source and hence composition of the oil and can vary between 6.0 and 8.0 barrels per tonne.
anticipates, the relative and absolute contribution of non-conventional liquids may increase. However, numerous technical, economic and environmental constraints make a rapid expansion of non-conventional production extremely challenging (Hirsch et al., 2005; Söderbergh et al., 2007). Hence, the supply of conventional oil in the period to 2030 is of critical importance.

The future supply of conventional oil will be shaped by multiple technical, economic and political factors, but the range of possibilities will be constrained by three physical features of the oil resource:

1. Production from individual fields normally rises to a peak or plateau, after which it declines as a result of falling pressure and/or the breakthrough of water (see Fig. 3 for example). While each field has a unique production profile as a result of both its physical characteristics and the manner in which it is developed, production normally begins to decline when less than half of the recoverable resources of the field have been produced (Höök et al., 2009b; IEA, 2008).

2. Most of the oil in a region tends to be located in a small number of large fields, with the balance being located in a much larger number of small fields (Sorrell et al., 2009). For example, there are some 70,000 oil fields in production worldwide, but approximately half of global production derives from only 110 fields, one quarter from only 20 fields and as much as one fifth from only 10 fields (IEA, 2008). Around 500 ‘giant’ fields account for around two thirds of all the crude oil that has ever been discovered.

3. These large fields tend to be discovered relatively early, in part because they occupy a larger area. Subsequent discoveries tend to be progressively smaller and often require more effort to locate. For example, over half of the world’s ‘giant’ fields were discovered more than 50 years ago (Robelius, 2007).

The effect of these physical features is illustrated by the production history of the UK Continental Shelf (Fig. 4). While the peak in production in the mid-1980s was mainly linked to the safety work that followed the Piper Alpha disaster, the subsequent peak in 1999 was largely ‘resource constrained’ in that the additional production from the small fields that were discovered relatively late became insufficient to compensate for the decline in production from the large fields that were discovered relatively early. This pattern was first observed for

---

Fig. 1. Classification of hydrocarbon liquids.
Note: The classification scheme used for liquid fuels is shown. No significance should be attributed to the relative size of each circle.

Fig. 2. Breakdown of 2007 liquid fuels production (mb/d).

Fig. 3. Production cycle of the Thistle field in the North Sea.
Source: UK Department of Energy and Climate Change.

Fig. 4. Oil production in the UKCS by field.
Source: Department of Energy and Climate Change.

---

3 There is no standard definition of the terms ‘peak’ or ‘plateau’, whether applied to individual field or an oil-producing region. For example, CERA (2008) define the end of plateau as when production from a field falls below 80% of the peak while Höök et al. (2009b) uses 95% and the IEA (2008) 85%. Multiple peaks are also possible for both individual fields (e.g. through the use of enhanced recovery methods) and oil-producing regions (e.g. due to political restrictions on production) (Sorrell and Speirs, 2009; Sorrell et al., 2009).

4 The collapse in oil prices in the late 1990s and the subsequent reduction in exploratory drilling was also a contributory factor, but the price rises in the first decade of the 21st century have not reversed the production decline.
individual US states and has since been repeated for over 100 large regions, including more than 60 countries around the world (Fig. 5). The same physical factors will ultimately drive a peak in global production, but given the multiple variables involved and the poor quality of available data, anticipating this peak is far from straightforward.

Global cumulative production of conventional oil stood at \( \frac{C}{24} \) 1128 billion barrels (Gb) in 2007, with an annual production of 29.5 Gb (\( \frac{C}{24} \) 80.7 mb/d) (IEA, 2008). Since 1995, global production has grown at an average of 1.5% per year (albeit with considerable variation from year to year), with 60% of cumulative production occurring since 1980 (Fig. 6). At current rates of production, the world uses as much conventional oil as the UK has ever produced (24 Gb) in only 10 months.

Oil reserves are the quantities of oil in known fields that are considered to be economically feasible to extract under defined conditions. Reserve estimates are inherently uncertain and are commonly quoted to two levels of confidence, namely proved (1P) and proved and probable (2P) reserves—although these terms are defined and interpreted in different ways. According to IHS Energy, global 2P reserves were approximately 1240 Gb at the end of 2007, or slightly more than cumulative production.

The sum of cumulative production and reserves is commonly referred to as cumulative discoveries. Cumulative discovery estimates will not be changed by production, but will be increased by the discovery of new fields and may be either increased or reduced by revisions to reserve estimates for known fields. The latter is commonly referred to as reserve growth since most estimates are revised upwards rather than downwards (cumulative discovery growth is a more accurate term). While some data sources record reserve revisions in the year in which they are made, others backdate the revisions to the year in which the relevant fields were discovered.
Most of the world’s conventional oil was discovered between 1946 and 1980 (Fig. 7) and since that time annual production has exceeded annual discoveries (Fig. 8). However, annual production has been less than official annual reserve additions over most of this period, since the latter also includes reserve growth at currently producing fields (Stark and Chew, 2005). Using data from IHS Energy, we estimate that between 2000 and 2007 an average of 48 Gb was added to global 2P reserves each year, split between 15 Gb/year of new discoveries and 33 Gb/year of reserve growth. However, these figures are uncertain and contested and many analysts expect the rate of reserve additions to decline.

3. Ultimately recoverable resource of conventional oil

Global forecasts of future supply of conventional oil vary in two basic respects, namely area beneath the curve and ‘shape’ of the curve. Since conventional oil is a finite resource, production forecasts must rise over time to a peak or plateau and then fall away. While multiple peaks are possible, a more plausible scenario is a multi-year plateau in which production fluctuates by a few percentage points before decline is established. The area under the curve, from when production begins to when it finally ends, represents the global ultimately recoverable resource (URR) of conventional oil. The URR is the sum of cumulative discoveries, future reserve growth at known fields and the volume of oil estimated to be economically recoverable from undiscovered fields—commonly termed the yet-to-find (YTF). The remaining recoverable resources are all the recoverable resources that have yet to be produced (Fig. 9). Global URR estimates are likely to increase over time as knowledge expands, prices increase and technology improves, but the rate of increase should slow as depletion advances. There is dispute over whether the impact at the global level will be significant or marginal. However, if technical or economic constraints mean that resources are accessible only in the long term, they will have little or no influence on the timing of peak production, although they could reduce the rate of post-peak decline.

Estimates of global URR for conventional oil vary widely in their methods, assumptions and results (Sorrell et al., 2009). Contemporary estimates fall within the range 2000–4300 Gb, while the corresponding estimates of the quantity of remaining recoverable resources fall within the range 870–3170 Gb. In other words, the highest estimate of remaining recoverable resources is nearly four times larger than the lowest estimate. The most in-depth assessment of the global URR is provided by the USGS (2000), whose mean estimate of 3345 Gb has been criticised by...
many as overly optimistic (Laherrère, 2001). This figure was recently updated by the IEA (2008) to 3577 Gb (Fig. 10). Aguilera et al. (2009) arrive at an even higher estimate of 4233 Gb, but some of their assumptions appear questionable (Sorrell et al., 2009). Also, while the rate of new discoveries appears to be lower than ‘anticipated’ by the USGS (Fig. 11), this is partly a consequence of restrictions on exploration in the most promising regions and failure to adjust discovery estimates to allow for future reserve growth. Nevertheless, even if the USGS/IEA estimates are broadly correct, it is likely that a large proportion of these resources (e.g. those in polar regions) will be accessed only relatively slowly, at high cost and with a declining net energy yield (Cleveland, 2005; Gagnon et al., 2009).

Although the USGS made simple and apparently optimistic assumptions about the rate of global reserve growth, these have subsequently proved consistent with observed trends (Klett et al., 2005; Sorrell et al., 2009). Also, while the rate of new discoveries appears to be lower than ‘anticipated’ by the USGS (Fig. 11), this is partly a consequence of restrictions on exploration in the most promising regions and failure to adjust discovery estimates to allow for future reserve growth. Nevertheless, even if the USGS/IEA estimates are broadly correct, it is likely that a large proportion of these resources (e.g. those in polar regions) will be accessed only relatively slowly, at high cost and with a declining net energy yield (Cleveland, 2005; Gagnon et al., 2009).

The implications of uncertainty in URR estimates for timing of the global peak can be explored with the help of a simple ‘bell-shaped’ model of regional oil production introduced by Hubbert (1956). With this model, the production cycle is symmetric and production peaks when half of the URR has been produced. Hubbert never claimed that regional production must follow a bell-shaped curve and empirical analysis suggests that it rarely does—more commonly, the rate of production decline is slower than the rate of increase (Brandt, 2007; Sorrell and Speirs, 2009). Nevertheless, the model provides a useful starting point. Using non-linear regression, we fitted the first differential of a logistic curve to global production data, constrained by different assumptions about the global URR (Fig. 12). For a URR estimate of 2500 Gb, this gives peak production in 2009 at a level of 82 mb/d, while for an estimate of 4500 Gb, it gives peak production in 2032 at a level of 115 mb/d. Hence, an 80% increase in the size of the URR (or a 250% increase in the size of the remaining resource), delays the date of peak production by only 23 years. Delaying the peak by 1 year requires addition of ~80 Gb to the global URR, which is two and half times greater than global production in 2007 and more than five times greater than global discoveries.

In practice, the global production cycle is more likely to be asymmetric. For example, production may increase more slowly prior to the peak as price signals provide incentives for demand reduction and as major producers restrict exports to maximise the size and longevity of revenues. Kaufmann and Shiers (2008) address these uncertainties by combining different assumptions about the global URR with different assumptions about the degree of asymmetry in the global production cycle. They constructed 64 possible scenarios for global production of conventional oil, 53 of which have a peak before 2031. Importantly, they find that large differences in the assumed URR or shape of the production cycle lead to relatively small differences in timing of peak production.

These simple calculations suggest that delaying the global peak in conventional oil production beyond 2030 requires a combination of a large URR (e.g. at least 3000 Gb), slow rates of production increase prior to the peak and/or a relatively steep decline in production after the peak. More rapid decline also results if the peak is extended into a multi-year plateau. However, lower rates of demand growth could delay the peak.

---

5 The USGS estimate refers to resources that had ‘…the potential to be added to reserves between 1995 and 2025 using existing technology’. This implies that the results could both underestimate the global URR (since some resources may be technically and economically accessible only in the longer term) and overestimate resource availability up to 2025 (since political and other constraints may prevent resources from being accessed and exploited).

6 This excludes ‘conventional oil produced by unconventional means’. The precise definition of this is unclear, but it includes extensive use of enhanced oil recovery techniques (EOR). The IEA (2008) provides a long-term oil supply cost curve that includes this category and suggests a larger URR of 4276 Gb. However, the IEA forecasts that only 12% (24 Gb) of the estimated potential from EOR will be produced before 2030.
4. Overview of contemporary global supply forecasts

This section summarises 14 contemporary forecasts of conventional oil supply in the period to 2030. Nine of these forecasts predict a peak in conventional oil production before 2030, while five do not. Each of these forecasts is produced by a different model and a wide range of approaches (or combinations of approaches) is used (Bentley et al., 2009).\textsuperscript{7} The models vary widely in the relative attention given to demand and supply, the attention paid to economic and physical variables, in the assumptions used for those variables and in the level of aggregation used for supply modelling (e.g. regional, country, field or project). Full descriptions of each of these models and forecasts are provided in Bentley et al. (2009), while their main features are summarised in Table 1. All forecasts were developed prior to the global economic recession in 2008 and hence do not reflect the subsequent reductions in upstream investment and oil demand.

Fig. 13 plots 13 of the global forecasts of all-oil while Table 2 compares estimates of cumulative production through to 2030 (BGR is omitted owing to insufficient data). In the few cases where alternative scenarios are provided only the ‘base case’ is shown. However, two scenarios are shown for Shell.

As Fig. 13 shows, the highest forecast production of all-oil in 2030 is over two-and-a-half times the lowest, while the range of forecast peak date ranges from the immediate past to the indefinite future. It may seem remarkable that detailed studies can reach such different conclusions. One cause for this range is that some of the forecasts exclude extra-heavy oil and for Miller and OPEC the contribution of NGLs has been estimated. However, this explains only a relatively small proportion of the difference.

One group of forecasts indicates an approximately linear growth in the production of all-oil to 2030, such that if the modellers foresee a peak it is beyond the end of their forecast (Fig. 14). These ‘quasi-linear’ forecasts are from the IEA, US EIA, OPEC and Exxon and all are driven primarily by demand modelling, with sources of supply being allocated to fill this demand. Meling’s model is superficially similar, reaching a peak in 2028, but the modelled supply fails to match the assumed demand from 2011 onwards.

The second group of forecasts indicates a peak before 2030, followed by a decline (Fig. 15). Shell is the only one of this group to explicitly model oil demand, and their all-oil peak is driven by declining demand from efficiency improvements and substitution rather than constrained supply. The remaining models use exogenous assumptions for demand growth and a combination of simple curve-fitting to regional trends in production and discovery, together with physical and economic modelling of the production from individual projects and fields. The forecasts from LBST, Campbell, Peak Oil Consulting, Uppsala, Total and Energyfiles initially follow their assumed demand trends before falling away. Meling’s forecast peaks late but does not meet the assumed 1.6%/year demand growth beyond 2011. Miller’s scenario depicts the maximum production that could be achieved were all ‘fallow fields’ (see below) and new discoveries to be developed immediately. Consequently this forecast shows an initial rise of potential capacity before falling away.

The annual rate of decline of all-oil production following the peak varies from 0.2% (Total) to 3.5–4% (LBST). The global URR

\textsuperscript{7} The information for each model was gathered from both published literature and the author(s), who commented on and approved written descriptions (Bentley et al., 2009). Several significant models are excluded (e.g. PFC Energy, Cambridge Energy Research Associates, World Energy Council), either because the individuals or organisations could not be contacted or because they were unable to collaborate for commercial or other reasons. However we do not consider that these omissions materially affect the conclusions.

### Table 1

<table>
<thead>
<tr>
<th>Category</th>
<th>Model</th>
<th>Liquids covered by model</th>
<th>Demand modelling?</th>
<th>Level of aggregation for supply modelling</th>
<th>URR (Gb)</th>
<th>Date of global peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>International</td>
<td>IEA</td>
<td>All-liquids</td>
<td>Yes</td>
<td>Physical/economic modelling of fields and projects. Regional curve-fitting for discovery trends</td>
<td>3577</td>
<td>No peak, but conventional oil plateau by 2030</td>
</tr>
<tr>
<td></td>
<td>OPEC</td>
<td>All-liquids</td>
<td>Yes</td>
<td>Mix of project and field (short term) and regional (long term). Model is largely demand-driven</td>
<td>3345</td>
<td>No peak</td>
</tr>
<tr>
<td>National</td>
<td>US EIA</td>
<td>All-liquids</td>
<td>Yes</td>
<td>Regional, with some individual country modelling</td>
<td>Not given</td>
<td>No peak</td>
</tr>
<tr>
<td></td>
<td>BGR</td>
<td>Conventional oil</td>
<td>No</td>
<td>Regional/global with simple mid-point peaking assumption</td>
<td>2979</td>
<td>~2020</td>
</tr>
<tr>
<td>Oil companies</td>
<td>Shell</td>
<td>All-liquids</td>
<td>Yes</td>
<td>Mix of project, field, basin and regional</td>
<td>Not given</td>
<td>~2030 (Blueprints)</td>
</tr>
<tr>
<td></td>
<td>Meling (StatoilHydro)</td>
<td>All-oil</td>
<td>No</td>
<td>Regional and country</td>
<td>3149</td>
<td>2028 (base case)</td>
</tr>
<tr>
<td></td>
<td>ExxonMobil</td>
<td>All-liquids</td>
<td>Yes</td>
<td>Mix of field, basin and country</td>
<td>Not given</td>
<td>2020</td>
</tr>
<tr>
<td>Consultancies</td>
<td>Energyfiles</td>
<td>All-oil</td>
<td>No</td>
<td>Mix of project, field, basin and country depending on data availability</td>
<td>2685</td>
<td>2017</td>
</tr>
<tr>
<td></td>
<td>LBST</td>
<td>All-oil</td>
<td>No</td>
<td>Field, country or regional. Simple curve-fitting for pre-peak countries</td>
<td>1840</td>
<td>2006</td>
</tr>
<tr>
<td>Universities and individuals</td>
<td>Peak Oil Consulting</td>
<td>All-oil, GTL and biofuels</td>
<td>No</td>
<td>Project and field level</td>
<td>Not estimated</td>
<td>2011–2013</td>
</tr>
<tr>
<td></td>
<td>Colin Campbell</td>
<td>All-oil</td>
<td>No</td>
<td>Regional and country. Assumes mid-point peaking and constant post-peak depletion rate</td>
<td>2425, all-oil; 1900, regular oil</td>
<td>2008</td>
</tr>
<tr>
<td></td>
<td>University of Uppsala</td>
<td>All-oil</td>
<td>No</td>
<td>Giants fields and country/regional for remainder</td>
<td>Not given</td>
<td>2008–2018</td>
</tr>
<tr>
<td></td>
<td>Richard Miller</td>
<td>Crude oil</td>
<td>No</td>
<td>Field (~3500 fields included)</td>
<td>2800</td>
<td>2013–2017</td>
</tr>
</tbody>
</table>
that is assumed or implied by these peaking forecasts ranges from 1840 Gb (LBST) to 3149 Gb (Meling).

Despite the inconsistencies in the coverage of liquids and the wide differences in methodology and assumptions, it is possible to compare these forecasts by focusing on a limited number of variables. The following section does this by examining the assumed or implied global URR of conventional oil and the interaction between the URR, the forecast date of peak production and the aggregate rate of production decline following the peak. The latter represents the net outcome of: (a) decline in production from fields that are past their peak; (b) production from fields that are ‘on plateau’ and (c) increase in production from new (and typically smaller) fields that are coming on-stream. Decline begins when (a) exceeds (c).

5. Framework for comparing global supply forecasts

Any forecast of production from a fixed resource can be divided into a growth phase and a decline phase, and perhaps a plateau phase. A change in height and date of the peak implies a change in either area under the curve or shape of the curve. The differences
between various forecasts presented here can therefore be viewed as differences in the assumed URR (i.e. the area under the curve), growth rate and/or decline rate and shape of the peak or plateau. Some of the forecasts use explicit assumptions for regional or global URR and/or shape of the production curve as input parameters, while most generate one or both as an output. But all can in principle be assessed according to how these parameters compare and whether they can be considered realistic.

Only eight of the models make explicit assumptions about global URR and these cover a wide range (Table 1). Those models that forecast a peak before 2030 estimate the URR of conventional oil to be in the range 1840–3150 Gb, while the three models that forecast no such peak and also provide estimates for URR suggest larger values of 3345–3577 Gb. Closer examination demonstrates that the models differ largely in their assumptions for reserve growth and YTF (Box 1). The smallest (mean) URR estimate is 1840 Gb (LBST) while the largest is 3577 Gb (IEA).

The interplay between URR, peak date and post-peak aggregate decline rate is shown in a simplified form in Fig. 16. Here, production is assumed to increase exponentially to a peak and then decrease exponentially at a different rate. This profile is unlikely in practice, but serves as a simple approximation. The other parameters have been fixed at values typical of the forecasts studied, specifically:

- Production continues for 100 years after peak, and the cumulative production by then is the effective URR. Fig. 16 shows URRs of 2600, 2800 and 3000 Gb.
- The growth rate to peak is 1.3%/year.
Box 1–Different assumptions about the components of global URR

- **Reserve** estimates range from 734 to 1332 Gb, but there are variations in definitions used and coverage of liquids and some models discount much of the reported Middle East reserves. The smallest estimate (734 Gb) is from Campbell, who also excludes extra-heavy oil, oil sands, deepwater fields, polar fields and NGLs. Excluding Campbell’s estimate, the difference between the largest and smallest estimates is 210 Gb.

- **Fallow fields** are fields that are discovered but not currently scheduled for development. The IEA (2008) estimates that 20% (257 Gb) of 2P reserves are located in 1874 fallow fields and they forecast cumulative production of 220 Gb from these fields by 2030. But some of these reserves may not be developed, since they are contained in small, isolated or complex fields that were discovered more than 20 years ago.

- **Reserves growth** is ignored by Campbell, who holds that industry databases provide accurate estimates of URR of individual fields. Some models, such as Energyfiles and Miller, make fairly conservative assumptions for reserve growth, while others implicitly accept USGS estimates, which were updated by the IEA (2008) to a mean value of 402 Gb.

- **Yet-to-find** estimates range from 114 to 805 Gb—a difference of some 690 Gb. Some models do not specify YTF directly, referring only to global URR. Forecasts relying on USGS estimates imply a YTF comparable to that assumed by the IEA (805 Gb).

The assumption of a 1.3%/year growth rate up to the peak is particularly important since (other things being equal) a faster rate of demand growth should lead to an earlier peak and vice versa. The 1.3% assumption is consistent with assumed or modelled growth rates in the majority of the forecasts, but these were developed prior to the global economic recession of 2008. The latter has reduced global oil demand, which could delay the rate of demand growth should lead to an earlier peak and vice versa.

The decline from peak is shown for values of 2%/year, 4%/year and 6%/year.

Production in 2007 is 85 mb/d and cumulative production by end 2007 is 1150 Gb.

The difference between production-weighted decline rate of post-peak fields and aggregate decline rate of total production needs to be met by incremental production from new projects. The volume of new resources that needs to be added each year will depend on the rate at which these resources can be produced – the ‘depletion rate’ – which is subject to physical, technical and economic constraints (Box 4). Taken together, these considerations constrain the minimum decline rate that can be considered reasonable. We consider that decline rates of less than 2.5%/year would be difficult to justify beyond 2030.

This analysis indicates that the two groups of forecasts differ largely in their assumed or implied URR, but the post-peak aggregate decline rate also plays an important role. Lower decline rates imply more optimistic assumptions for global URR (e.g. 3000–3600 Gb for 3%/year), but if the latter is set to more conservative levels the required decline rate appears both inconsistent with current evidence and disturbingly high in terms of its likely effects on society. For example, a 6%/year aggregate decline rate implies loss of two thirds of conventional oil production within 20 years.

Hence, one way of comparing global supply forecasts is to identify the assumed or implied global URR and aggregate decline rate and to assess the plausibility of these assumptions. Some priority should therefore be given to obtaining the data that constrain these parameters. We anticipate an improvement in the understanding of decline rates at every scale, building on the recent studies by the IEA (2008), Höök et al. (2009a, 2009b) and CERA (2008). Unfortunately, a consensus on global URR appears much less likely in the near term.
6. Impacts of rates of discovery, reserves growth and depletion on date of peak

An alternative approach to judging the plausibility of the forecasts is to ask how realistic is the assumed or implied URR for conventional oil and date of peak given the historical trends in both new discoveries and reserve growth. Here we present some simple ways of looking at this issue.

One approach is to apply Hubbert’s original rule of thumb that production in a region peaks when half of the URR has been produced. Using the URR estimates from the USGS (2000), we can investigate this rule by estimating depletion at peak for 37 countries that have passed their peak production (Fig. 5).8 This gives a simple mean of 22%, a production-weighted mean of 26% and a maximum of 52%. In other words, most countries appear to

---

**Box 2—Assumed or implied URR, decline rate and date of peak for the peaking forecasts**

- Campbell forecasts a peak in 2008 with an all-oil URR of 2450 Gb. The post-peak aggregate decline rate is ~2%/year, which is the lowest of all forecasts except for Total and is one reason for the early peak.
- Peak Oil Consulting focuses on near-term production up to 2016, where supply of oil from identified new projects is offset by decline in production from existing fields. Near-term supply is tightly constrained by lead time of major projects, because those that will come on-stream within this period must already be committed.
- Energyfiles forecasts a peak in 2017 with a URR of 2685 Gb.
- Miller’s model estimates the maximum production that could be achieved, regardless of cost or demand. The URR is 2800 Gb and the peak is around 2018. In practice, the potential excess before 2018 caused by the bank of fallow fields is likely to be deferred, leading to a later peak.
- The BGR forecast has very little information except a URR of just under 3000 Gb and a peak in 2020. The implied decline rate must therefore be about 2.5%/year.
- Total forecasts an all-oil peak at 2020. Although decline rate and URR are not stated, we calculate post-peak aggregate decline to be 0.2%/year This implies a URR of at least 4500 Gb, which may be consistent with Total’s assumptions. Alternatively, the aggregate decline rate may steepen after 2030.
- Uppsala University forecast a peak between 2008 and 2018. They do not state a URR and implicitly assume that the YTF will have little effect on date of the peak. Their forecast indicates an initial rapid aggregate decline, which finally levels off to just under 2%/year.
- Meling’s forecast has a peak in 2028 and a URR of about 3150 Gb.
- LBST assumes a URR of 1840 Gb and estimates that the peak has already occurred. The aggregate post-peak decline rate is between 3.5%/year and 4%/year.
- Shell forecasts a demand-driven peak for all-oil production around 94 mb/d in 2030 in the Blueprints scenario, and around 91 mb/d in 2020 in the Scramble scenario.

---

8 Since the USGS estimates reserve growth only at the global level, this was allocated between countries in proportion to their estimated URR excluding reserve growth. Post-peak countries for which URR estimates were not available were excluded, as was Russia. It is important to note that timing of the peak production for many of these countries has been influenced by factors other than physical depletion.
have reached their peak of conventional oil production well before 50% of their (USGS estimate of) URR has been produced. If we take the IEA (2008) mean URR estimate of 3577 Gb and assume production growth of 1.3%/year, the date of ‘mid-point’ peak is 2027. Larger resources and/or a slower rate of production growth would delay the mid-point while smaller resources and/or more rapid growth would bring it forward. With the possible exception of Total, none of the forecasts assumes or implies a larger mean estimate of URR and most assume production growth of around 1.3%/year. Since historical experience suggests that the mid-point rule is optimistic, this simple calculation sounds a cautionary note to the quasi-linear forecasts. A second approach is to apply a rule of thumb proposed by PFC Energy that production peaks in most regions when cumulative discoveries reach 60% of cumulative 2P discoveries. We can investigate this by estimating the ratio of cumulative production to cumulative discoveries at peak for 54 post-peak countries. This gives a simple mean of 38% and a production-weighted mean of 36%. In other words, most countries to date appear to have reached their peak production well before 60% of their cumulative discoveries have been produced.

To apply the PFC rule at the global level, we need an estimate of future rate of reserve additions through new discoveries and reserve growth. Using the IHS Energy database, the current rate can be estimated as ~48 Gb/year since 2007, split between 15 Gb/year of new discoveries and 33 Gb/year of reserve growth (Sorrell et al., 2009). In 2007, the global ratio of cumulative production to cumulative discoveries was approximately 48%. Under the optimistic assumption that rate of discovery remains unchanged, the expected date of peak production can be estimated from ‘the PFC rule’ using different assumptions for demand growth and contribution from reserve growth (Table 3). These calculations suggest that to postpone the peak beyond 2030 requires sustaining annual reserve additions around 35 Gb/year, even for low rates of annual demand growth (1%/year). It also implies a URR of at least 3200 Gb. To do the same with higher rates of demand growth requires correspondingly higher rates of annual reserve additions and a larger URR. But discovery is on a long-term declining trend (Fig. 7) and as global production shifts towards newer, smaller and offshore fields the rate of reserve growth is expected to decrease in both percentage and absolute terms (Grace, 2007; Sorrell et al., 2009; Verma and Ulimshek, 2003). Since historical experience suggests that the ‘PFC rule’ may be optimistic, this also sounds a cautionary note to the quasi-linear forecasts. A third approach is to investigate the depletion rates implied by the different forecasts and to assess whether these are consistent with historical experience (Box 4). At the regional level maximum depletion rates have typically been around 3%/year, with the most rapid depletion occurring in offshore basins (Aleklett et al., 2009; Höök, 2009; Jakobsson et al., 2009). The IEA (2008) is the only quasi-linear forecast that provides sufficient detail to allow depletion rates

(footnote continued)

(although uncertain) estimate for global average recovery factor of 35% gives an estimated original-oil-in-place for discovered fields of 7540 Gb. Assuming that an average of 50% recovery can be achieved in 50 years leads to an estimated annual reserves growth from discovered fields of 22 Gb/year and URR from these fields of 3770 Gb compared with the 2771 Gb assumed by the IEA. A better approach would be to assess the recovery potential of individual fields, but to our knowledge, no such study has been attempted at the global level.
to be estimated for different categories of field. The IEA projects 114 Gb of new discoveries by 2030 (5 Gb/year) and forecasts these fields producing 19 mb/d by 2030 with cumulative production of 46 Gb (Aleklett et al., 2009). This implies an average depletion rate of 10% (and rising) from these resources by 2030, which far exceeds the historical experience of any major oil-producing region. A comparable analysis for fallow fields suggests even higher average depletion rates of 12–13% (and rising) by 2030. In addition to physical constraints, it is questionable whether OPEC producers would have the incentive to deplete these resources at the rates implied (Déès et al., 2007; Gately, 2004). But the use of more realistic depletion rates would either imply an earlier peak or require higher rates of discovery and reserve growth. By implication, the same conclusions are likely to apply to the other quasi-linear forecasts.

The above rules are approximate and should be used with considerable caution. A better approach is to model production expected from existing and yet-to-find fields under realistic assumptions for production cycles, discovery rates and likely recovery gains from improved technology. Several of the bottom-up models do this, although their specific assumptions need to be made more transparent and the sensitivity to those assumptions more carefully explored. Nevertheless, the fact that all these models forecast a peak occurring before 2030 provides a further note of caution to the quasi-linear results.

7. Discussion

Comparison of forecasts is not straightforward owing to the lack of transparency of most models, inconsistencies in the definition and coverage of different liquids and the wide range of methods and assumptions used. With the exception of Shell, all models are used for ‘single-value’ forecasts and none has been employed to explore a wider range of plausible socioeconomic scenarios. Comprehensive sensitivity testing is essential given the degree of uncertainty over many relevant variables, but apart from high and low oil price assumptions it remains the exception rather than the rule.

Nevertheless, contrary to initial impressions, there is a degree of convergence appearing in the forecasts reviewed. Although the range of modelling approaches is wide, the range of dates for the final peak can be closely linked to different explicit or implicit assumptions for the URR and/or post-peak global aggregate production decline rate. Other differences are arguably either secondary or are components of these two parameters. Even some of the quasi-linear models now foresee a levelling off of conventional oil production around 2030. We anticipate that this convergence will continue as more accurate data become available, leading to a greater consensus on plausible oil supply futures.

While the above analysis excludes non-conventional liquids, uncertainties for policy makers concern not only the timing of peak of conventional oil but also availability and cost of substitute fuels. Given the speed and scale of substitution that is likely to be sought, both economic potential of non-conventional liquids and the associated lead times deserve urgent and careful study.

For short-term forecasting, the approach of groups such as Peak Oil Consulting appears fairly robust given the lead times on major new projects. However, none of the forecasts reviewed takes into account the impact of global economic recession. If demand recovers over the next 2–4 years, constraints on oil supply may well occur as a result of delayed upstream investment (IEA, 2009).

For medium- to long-term forecasting, the number and scale of uncertainties multiply, making precise forecasts of the date of peak production unwarranted. Nevertheless, it is possible to form
some judgment of plausibility of forecasts that delay the peak of conventional oil production until after 2030. These would appear to require some combination (although not all) of the following conditions:

- lower rates of demand growth than are assumed in the forecasts reviewed here (e.g. 1%/year);
- a global URR that is comparable to or greater than the mean estimate of the USGS (~3600 Gb);
- a rapid decline in production following the peak (e.g. 3%/year or more);
- cumulative production at the date of peak that exceeds 50% of the global URR (i.e. much greater than previously observed in the majority of post-peak regions);
- cumulative production at the date of peak that exceeds 60% of cumulative 2P discoveries (i.e. again much greater than previously observed in the majority of post-peak regions);
- an annual rate of new discoveries over the period to 2030 that equals or exceeds that achieved over the last decade (i.e. reversing the trend of the last 40 years, despite the declining size of newly discovered fields);
- an annual rate of reserve growth over the period to 2030 that equals or exceeds that achieved over the last decade (despite the growing share of new, smaller and offshore fields that have less potential for reserve growth);
- depletion of these resources at an average rate that is several times greater than the maximum rate previously achieved in any oil-producing region; and
- favourable ‘above-ground’ conditions, including appropriate incentives for investment, sufficient access to prospective areas, political stability and so on in all major oil-producing regions.

‘Quasi-linear’ forecasts need to either demonstrate how such conditions can be met or why they do not apply. In our judgment, most of these conditions appear optimistic and more so in combination. This, together with the analysis in Section 3, leads us to the conclusion that a peak of conventional oil production before 2030 appears likely.

Forming a judgement on the timing of peak production in the interim is more difficult (and less important than acknowledging that it is likely and taking appropriate mitigating actions). Sorrell et al. (2009) conclude that larger estimates for global URR may be reasonable and hence the assumptions of some of the ‘peaking’ forecasts overly pessimistic. However, a mix of above- and below-ground constraints is likely to prevent these resources from being developed at the rate that is implied by the quasi-linear forecasts. Also, most models do not capture the complex interactions between supply and demand, the likely consequence of which is to turn a sharp peak into a bumpy plateau. The current economic recession may delay the point at which physical depletion leads to supply constraints, but if demand recovers relatively soon the difference is unlikely to be more than a few years. For similar reasons, climate policy seems unlikely to have a significant impact in the medium term, given the anticipated growth in oil demand in Asia and Middle East. The ability of the oil market to signal when, how quickly, and market driven. Ecological Economics 67 (3), 405–411.

Given these complexities, we suggest that there is a significant risk of a peak in conventional oil production before 2020. At present, most OECD governments are failing to give serious consideration to this risk, despite its potentially far-reaching consequences.

Acknowledgements

This paper is based on an in-depth comparison of global supply forecasts by Bentley et al. (2009), which in turn formed part of a wide-ranging study of global oil depletion by the UK Energy Research Centre (Sorrell et al., 2009). The authors would like to thank the UK Research Councils for their financial support, IHS Energy for allowing publication of data from their PEPS database, and the following individuals for providing model descriptions: Michael Smith (Energyfiles), Jörg Schindler and Werner Zittel (Ludwig Bölkow Systemtechnik), Chris Skrebowskio (Peak Oil Consulting), Leif Magne Meling (StatollHydro), Laurent Maurel (Total Exploration and Supply), Colin Campbell (ASPO), John Staub (US Energy Information Administration), Garry Brennand (OPEC) and David Freedman (Shell E&P). The usual disclaimers apply.

References

Hubbert’s Legacy: A Review of Curve-Fitting Methods to Estimate Ultimately Recoverable Resources

Steve Sorrell1,3 and Jamie Speirs2

A growing number of commentators are forecasting a near-term peak and subsequent terminal decline in the global production of conventional oil as a result of the physical depletion of the resource. These forecasts frequently rely on the estimates of the ultimately recoverable resources (URR) of different regions, obtained through the use of curve-fitting to historical trends in discovery or production. Curve-fitting was originally pioneered by M. King Hubbert in the context of an earlier debate about the future of US oil production. However, despite their widespread use, curve-fitting techniques remain the subject of considerable controversy. This article classifies and explains these techniques and identifies both their relative suitability in different circumstances and the level of confidence that may be placed in their results. This article discusses the interpretation and importance of the URR estimates, indicates the relationship between curve fitting and other methods of estimating the URR and classifies the techniques into three groups. It then investigates each group in turn, indicating their historical origins, contemporary application and major strengths and weaknesses. The article then uses illustrative data from a number of oil-producing regions to assess whether these techniques produce consistent results as well as highlight some of the statistical issues raised and suggesting how they may be addressed. The article concludes that the applicability of curve-fitting techniques is more limited than adherents claim and that the confidence bounds on the results are wider than usually assumed.

KEY WORDS: Petroleum resource estimation, peak oil.

INTRODUCTION

A growing number of commentators are forecasting a near-term peak and subsequent terminal decline in the global production of conventional oil as a result of the physical depletion of the resource (Campbell and Heapes, 2008; Aleklett and others, 2009). These forecasts frequently rely on the estimates of the ultimately recoverable resources (URR) of different regions, obtained through the use of curve-fitting techniques. Curve-fitting was originally pioneered by M. King Hubbert in the context of an earlier debate about the future of US oil production (Bowden, 1985). However, despite the widespread and increasing use of these techniques, they remain the subject of considerable controversy.

In a seminal article, Hubbert (1956) forecast the future of US oil supply by fitting a curve to historical production and projecting this forward under

1Sussex Energy Group, SPRU (Science & Technology Policy Research), Freeman Centre, University of Sussex, Falmer, Brighton BN1 9QE, UK.
2Imperial College Centre for Environmental Policy and Technology (ICCEPT), Imperial College London, South Kensington Campus, Exhibition Road, London SW7 2AZ, UK.
3To whom correspondence should be addressed; e-mail: s.r.sorrell@sussex.ac.uk
4These involve selecting and parameterising a mathematical function that has the ‘best’ statistical fit to a set of data points, possibly subject to constraints.
5Hubbert’s original curve was hand drawn, but his later articles used more formal techniques.
the assumption that, first, production must eventually decline exponentially; and second, the area under the curve must equal the URR of the United States. Using the industry consensus estimates for the US URR (150–200 billion barrels), Hubbert forecast that the US oil production would peak sometime between 1965 and 1971. Many commentators considered Hubbert’s approach to have been vindicated when the US production peaked in 1970 and began its long decline (Strahan, 2007). However, subsequent analysis has shown that the accuracy of his forecast was partly fortuitous (Kaufmann and Cleveland, 2001; Cavallo, 2005a, b).

Shortly after the publication of Hubbert’s article, the consensus on the US URR estimates evaporated. Numerous commentators disputed Hubbert’s approach, and a series of more optimistic estimates of the US URR began to appear. Most famously, the US Geological Survey (USGS) estimated a value of 580 billion barrels (Gb) for the lower 48 states, based on forecasts of future drilling activity (Zapp, 1962). Such discrepancies motivated Hubbert to develop more formal methods to estimate the URR based on curve-fitting to production and discovery data. Hubbert developed and applied several such techniques between 1962 and 1982, all of which produced estimates in the range 150–200 Gb. These techniques have since been adopted and developed by numerous authors (Laherrère, 2000a, b, 2002a, b, 2004a, b; Campbell and Heapes, 2008; Mohr and Evans 2008), and the results have been used to underpin forecasts of near-term peaks in regional and global oil production. Although dismissed by critics as ‘trendology’ (Charpentier, 2003; Lynch, 2003), curve-fitting techniques have much in common with more sophisticated methods of estimating resource size such as discovery process modelling. They may also have an important role to play when, as generally the case, only aggregate data on production or discovery in a region is available—as opposed to detailed geological information or data on individual fields.

This article classifies and explains these curve-fitting methods and identifies both their relative suitability in different circumstances and the level of confidence that may be placed in their results. The article is based on a comprehensive review by Sorrell and Speirs (2009). The article begins by discussing the interpretation and importance of the URR estimates, indicating the relationship between curve-fitting and other methods for estimating URR and classifying the curve-fitting techniques into three groups. The subsequent sections investigate each group in turn, indicating their historical origins, contemporary application, and major strengths and weaknesses. The article then uses illustrative data from a number of oil-producing regions to assess whether these techniques produce consistent results, highlights some of the statistical issues raised and suggests how they may be addressed. The article concludes that the applicability of curve-fitting techniques is more limited than many adherents claim and that the confidence bounds on the results are wider than those that are usually assumed.

METHODS FOR ESTIMATING ULTIMATELY RECOVERABLE RESOURCES

The ultimately recoverable resources (URR) of a region represent the amount of resources that are anticipated to be recovered from when production begins to when it finally ends—although for practical purposes, estimates are usually made for a shorter period of time (e.g. up to 2050). Estimates of the URR require specification of the boundaries of the relevant region, the hydrocarbons covered, the timeframe for which the estimate is made, the relevant technical and economic assumptions and the associated range of uncertainty. Lack of clarity over such issues, together with the tendency to produce single value estimates, explains much of the controversy over this topic (Rogner, 1997). While not always treated as such, the URR is an endogenous variable. For example, increasing prices should make marginal resources more profitable as well as inducing technical improvements that reduce production costs, improve recovery factors and allow access to previously inaccessible resources. At issue is whether such developments can be expected to significantly increase the recoverable resources from a region, or whether only marginal increases may be expected. This in turn will depend on the characteristics of the region, the extent to which it has been explored, the technology currently used and the timeframe under consideration.

Estimates of the URR may be developed for levels of aggregation ranging from individual fields to the whole world, with different techniques being more or less suitable for different levels. Aggregate estimates may be derived from the estimates developed at lower levels (e.g. for individual fields), but the latter can only be summed arithmetically if they represent mean estimates (Pike, 2006). Estimates
are frequently made at the country or regional level, but these typically encompass several geologically distinct areas that may also extend into neighbouring countries or regions. The lack of geological homogeneity within such boundaries can greatly complicate resource assessments when only aggregate data is used (Charpentier, 2003).

There are a variety of methods for estimating the URR and many variations on the basic techniques. The appropriate choice depends on the nature of the region under study, and the data and human resources available. The most reliable estimates are likely to be derived from a combination of methods, and evidence suggests that the results can vary widely (Divi, 2004; Ahlbrandt and Klett, 2005). Most of the methods associated with Hubbert may be characterised as producing single-value estimates from the extrapolation of curves fitted to historic data on production or discoveries for aggregate regions such as an oil-producing country. There is relatively little use of geological or other information, and the methods are simple to apply using data that is either available in the public domain or available at reasonable cost from commercial sources such as IHS Energy. In contrast, the methods used by the USGS and others produce probabilistic estimates from geological assessments of disaggregate regions, with extensive use of geological information and statistical techniques (USGS, 2000). These methods are complex and resource intensive and rely on extensive data sources that are usually inaccessible to third parties. This characterisation is an oversimplification, however, as there are considerable overlaps between the two, especially for regions that are at a relatively mature stage of exploration and production (Drew and Schuenemeyer, 1993).

For the less-explored regions, estimates must rely on the geological analysis of seismic and other data. A traditional approach, usually applied at the basin level, is to estimate hydrocarbon volumes by multiplying the estimated sedimentary volume by an estimated yield in barrels per cubic kilometre (Weeks, 1952; White and Gehman, 1979; Gautier, 2004). For unexplored areas, the values for such calculations are based on measurements from geologically similar regions where more information is available. Other approaches are applied at lower levels of aggregation and typically use Monte Carlo methods to multiply estimates or measurements of variables such as pore volume, porosity and oil saturation. Data may only be available for a subset of these variables and, for unexplored areas (e.g. the Arctic), such estimates must necessarily have large confidence bounds. In Hubbert’s view:

“…it is easy to show that no geological information exists, other than that provided by drilling, that will permit an estimate to be made of the recoverable oil obtainable from a primary area that has a range of uncertainty of less than several orders of magnitude.” (Hubbert, 1982)

Estimates may also be made by combining the expert judgment of several geologists (Baxter and others, 1978). Typically, each geologist reviews the relevant information and then estimates either a single value or a probability distribution for each of the relevant factors which are then combined into a probability distribution that reflects the full range of opinions (White, 1981; Gautier, 2004). This method is appropriate for all the levels of aggregation and data availability, and can accommodate exploration constraints and other factors that may be poorly handled by other methods (Charpentier and others, 1995). However, it lacks transparency and relies heavily on the knowledge and objectivity of the individual assessors.

For well-explored regions, more reliable estimates can be obtained by using data on discovered fields. There are three approaches:

- **Field-size distributions**: Estimates of the URR may be derived by combining data on the sample of discovered fields with assumptions about the size distribution of the underlying population of fields. For example, if the field population is assumed to take a ‘power-law’ distribution, a plot of the number of fields exceeding a particular size on logarithmic scales should approximate a straight line. In this case, undiscovered resources may be estimated by plotting a cumulative frequency distribution on a log scale, fitting a linear regression, extrapolating this to smaller field sizes and calculating the area under the curve (Cramer Barton and La Pointe, 1995). This estimate is sensitive to the point at which the observed size distribution is curtailed when fitting the curve, as well as to assumptions about the minimum viable field size and the appropriate functional form for the population size distribution (which is an enduring focus of controversy) (Charpentier and others, 1995; Laherrère, 2000a, b; Kaufman, 2005). An alternative approach plots cumulative discoveries as a function of the rank of
the field (where the largest field is rank 1) and estimates the URR from the asymptote to which the curve is trending. This approach has much in common with the ‘discovery projection’ technique described later.6

- **Discovery process modelling:** The discovery projection techniques involve statistical analyses of the number and size of discovered fields as a function of either time, the discovery sequence or some measure of exploratory effort, such as the number of exploratory wells drilled (Arps and Roberts, 1958; Kaufman, 1975; Meisner and Demirmen, 1981; Forman and Hinde, 1985). This technique is sometimes combined with assumptions about the field size distribution and/or information about field location (Schuenemeyer and Drew, 1994). Several of these models simulate a probabilistic law governing the process of new field discovery and can be used to provide forecasts of the number, size and sequence of future discoveries together with the anticipated success rate of exploratory drilling (Power, 1992; Power and Fuller, 1992a, b).

- **Curve-fitting:** The curve-fitting methods use regression techniques to fit curves to the historic trends in discovery or production in a region and extrapolate the curves to estimate the URR. The explained variables may be cumulative production, the rate of production, cumulative discoveries or the rate of discoveries, while the explanatory variable may either be time or some measure of exploratory effort, such as the cumulative number of exploratory wells. A major advantage of curve-fitting techniques is that they do not require data on individual fields—which from many regions of the world is unavailable, unreliable or very expensive to obtain. Curve-fitting was pioneered by Hubbert (1956, 1959, 1962, 1982) and has subsequently been adopted and developed by numerous analysts, including in particular those concerned about ‘peak oil’ (Cleveland and Kaufmann, 1991; Campbell, 2002; Laherrère, 2003; Imam and others, 2004; Mohr and Evans 2008).

---

6If the discovery rate was constant and if fields were discovered precisely in the descending order of size, the field-rank and discovery projection techniques would be identical. But the advantage of discovery projection is that it does not require data on individual fields.

---

The above three ‘extrapolation’ techniques all assume that: the field size distribution is highly skewed, with the majority of oil being located in a small number of large fields; and that the large fields tend to be discovered relatively early, with subsequent discoveries being progressively smaller and often the product of increasingly greater effort. These features should be reflected in both the size distribution of the discovered fields and the ‘shape’ of the discovery or production cycle. For example, Figure 1 plots cumulative discoveries in a region as a function of time. As the average size of the newly discovered fields falls, the curve trends towards an asymptote which can be taken as an estimate of the regional URR. All of the curve-fitting techniques rely on patterns such as this.

The extrapolation techniques are best applied to geologically homogeneous areas that have had a relatively unrestricted exploration history (e.g. without areas being closed to exploration for legal or political reasons). If this is not done, then the mixing of different populations of fields or the opening up of new areas for exploration (e.g. new plays within a basin) can lead to inconsistencies in the time-series and undermine the basis for estimating size distributions and extrapolating historical trends (Wendebourg and Lamiriaux, 2002). However, even if the region is geologically homogeneous, various technical, economic and political factors (e.g. the advent of horizontal drilling) can lead to structural breaks in a time-series (Harris, 1977).7 Since all the historical

---

7Exploration is very rarely unrestricted in aggregate regions. For example, the US imposes far fewer restrictions than most countries, but the Arctic National Wildlife Refuge, the eastern Gulf of Mexico, much of the western offshore and many onshore areas in the Rockies are off-limits for environmental reasons (Mills, 2008).
trends reflect the net effect of physical depletion, technical change and numerous economic, political and institutional influences, the extrapolation of those trends can only be expected to provide reliable URR estimates if physical depletion provides the dominant influence and continues to do so in the future. This condition will not apply in all oil-producing regions although depletion should become increasingly important as exploration proceeds.

Curve-fitting techniques are very popular owing to their simplicity and the relative availability of the relevant data. These techniques estimate the URR for a region by extrapolating historical trends in aggregate data, but they vary in their choice and definition of the explained and explanatory variables (Table 1). Other explanatory variables can and should be included in the specification, but generally are not. The following sections discuss each group of techniques in turn.

**PRODUCTION OVER TIME TECHNIQUES**

The simplest, although not the most reliable method of estimating the URR uses non-linear regression\(^8\) to fit a curve to time-series data on cumulative production. This curve may take a variety of forms with its shape being defined by three or more parameters, one of which corresponds to the URR. Hubbert (1982) assumed a *logistic model* (Box 1) which implies that cumulative production will initially grow exponentially, but the rate of growth will fall and eventually decline to zero as the URR is approached.\(^9\) The curve is defined by three parameters, representing the URR, the steepness of the curve and the midpoint of the growth trajectory, with the production cycle being obtained from the first differential of this curve. The URR may also be estimated by fitting a curve to the production data, although the result may well be different (Carlson, 2007). While a ‘bell-shaped’ production cycle is usually referred to as a ‘Hubbert curve’, Hubbert repeatedly states that it need not take this form (Hubbert, 1982).

Curve-fitting to production trends should be more reliable if production has passed its peak and is only viable if the rate of increase of production has passed its peak (i.e. the point of inflection on the rising production trend). The US data fits the logistic model relatively well (Fig. 2) despite covering a period that includes two world wars, several recessions, two oil shocks, revolutionary changes in

---

**Table 1. Classification of curve-fitting techniques by their choice of explained and explanatory variables**

<table>
<thead>
<tr>
<th>Group</th>
<th>Technique</th>
<th>Explained variable</th>
<th>Explanatory variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production over time</td>
<td>Cumulative production projection</td>
<td>Cumulative production</td>
<td>Time</td>
</tr>
<tr>
<td></td>
<td>Production projection</td>
<td>Rate of production</td>
<td>Time</td>
</tr>
<tr>
<td></td>
<td>Production decline curve</td>
<td>Rate of production</td>
<td>Cumulative production</td>
</tr>
<tr>
<td>Discovery over time</td>
<td>Cumulative discovery projection</td>
<td>Cumulative discovery</td>
<td>Time</td>
</tr>
<tr>
<td></td>
<td>Discovery projection</td>
<td>Rate of discovery</td>
<td>Time</td>
</tr>
<tr>
<td></td>
<td>Discovery decline curve (time)</td>
<td>Rate of discovery</td>
<td>Cumulative discovery</td>
</tr>
<tr>
<td>Discovery over exploratory effort</td>
<td>Creaming curve</td>
<td>Cumulative discovery</td>
<td>Exploratory effort</td>
</tr>
<tr>
<td></td>
<td>Yield per effort curve</td>
<td>Rate of discovery wrt exploratory effort</td>
<td>Exploratory effort</td>
</tr>
<tr>
<td></td>
<td>Discovery decline curve (effort)</td>
<td>Rate of discovery wrt exploratory effort</td>
<td>Cumulative discovery wrt exploratory effort</td>
</tr>
</tbody>
</table>

Notes: (1) The terms used to label these techniques are not standardised. (2) Rate of production is the first derivative of cumulative production with respect to time. Alternative terms are the rate of change of cumulative production, or more simply production. Similar comments apply to the rate of discovery, although here the derivative may be with respect to either time or exploratory effort. See Sorrell and Speirs (2009) for mathematical definitions of the explained and explanatory variables.

---

\(^8\)Non-linear regression is straightforward with modern computer technology, but the earlier literature uses simpler methods such as the linear transformation of the functional form followed by a linear regression (Hubbert, 1982). If the production or discovery cycle is well advanced, then it is possible to estimate the URR through visual identification of the asymptote to which the curve is trending (Bentley, 2009).

\(^9\)Hubbert (1982) begins with an assumed parabolic relationship between production and cumulative production and uses this to derive a logistic equation for cumulative production over time. However, this formal derivation came more than 20 years after he first referred to the logistic model (Hubbert, 1959; Sorrell and Speirs, 2009).
technology and the opening up of new oil-producing regions. In contrast, the logistic model provides a relatively poor approximation to the production cycle for most other oil-producing regions, especially when this occurs over a shorter period of time (Brandt, 2007).

The logistic is one of a family of symmetric and asymmetric curves that are widely used to model growth processes (Meade, 1984; Tsoularis and Wallace, 2002). Alternatives include the generalised logistic (Nelder, 1971), Bass (1969), Gompertz (Moore, 1966) and bi-logistic (Meyer, 1994) as well as the cumulative lognormal, Cauchy and Weibull distributions (Wiorkowski, 1981; Meade, 1984). One of the few that has been applied to oil depletion is the cumulative normal, which Bartlett (2000) fitted to the US production data. While there is no robust theoretical basis for choosing between these models, there are a variety of reasons for expecting the production cycle to be asymmetric. For example, production could decline rapidly after the peak as the large fields are depleted, or could decline more slowly if enhanced oil recovery (EOR) techniques are used. Brandt (2007) analysed 74 oil-producing regions and found that the rate of production increase exceeded the rate of decline in over 90% of
cases, suggesting that an asymmetric model may be more appropriate. However, when Moore (1962) fitted an asymmetric Gompertz function to the US data he obtained an URR estimate that was almost twice as large as that from the logistic model for a comparable goodness of fit. Very similar results were later obtained by Wiorowski (1981) and Cleveland and Kaufmann (1991). This highlights a generic weakness of curve-fitting techniques, namely: different functional forms often fit the data comparably well but give very different estimates of the URR (Ryan, 1966).

Production cycles often have more than one peak as a result of economic, technical or political changes or the opening up of a new region (Laherrère, 2000a, b). For example, Illinois experienced two

10Wiorowski (1981) compared a ‘Generalized Richards’ model (which can take an exponential, logistic, or Gompertz form depending on the parameters chosen) with a cumulative Weibull and found that they fit the US cumulative production data equally well but led to significantly different URR estimates (445 and 235 Gb, respectively).

11Cleveland and Kaufmann (1991) fitted a logistic curve to the US production data through to 1988 and found that the adjusted $R^2$ changed only from 0.9880 to 0.9909 as the value of the URR varied from 160 to 250 billion barrels.

If cumulative production grows logistically, then a plot of the ratio of production to cumulative production as a function of production should be approximately linear (Hubbert, 1982). If a linear regression is fit to this data, then the URR may be estimated by extrapolating and identifying the intersection with the cumulative production axis (Fig. 3). This straightforward technique was popularised by Deffeyes (2005) and is sometimes termed ‘Hubbert Linearisation’ (HL). While methodologically
straightforward, it is equivalent to fitting a logistic curve to cumulative production and hence will be unreliable if (as is usually the case) cumulative production departs from the logistic model.

The linearisation technique is closely related to the *decline curve analysis* used by reservoir engineers to project the future production of individual wells or fields (Arps, 1945; Arps, 1956; Gowdy and Roxana, 2007). If production declines exponentially, then the URR for the field may be estimated by plotting production against cumulative production, fitting a linear regression and extrapolating this until it crosses the cumulative production axis (Fig. 4). This technique is widely used, but alternative functional forms for decline curves should also be investigated since the exponential model can underestimate the URR (Kemp and Kasim, 2005; Li and Horne, 2007). Nevertheless, a ‘production decline curve’ for an individual field may be expected to be lead to more reliable estimates of URR than an equivalent technique for an oil-producing region.

In sum, curve-fitting to production trends is straightforward and relies on data that are readily available, relatively accurate and free from the complications of reserve growth (see below). However, while these techniques may sometimes provide reliable estimates in regions that are well past their peak of production, they have important drawbacks including the lack of a robust basis for the choice of functional form; the sensitivity of the estimates to that choice; the risk of ‘over-fitting’ multi-cycle models; the inability to anticipate future production cycles; and the neglect of economic, political and other variables that have shaped and will continue to shape the production cycle. These drawbacks are also shared by the discovery-based techniques described shortly.

**DISCOVERY OVER TIME TECHNIQUES**

Curve-fitting to discovery trends was first introduced by Hubbert (1962, 1966, 1982) and has since been employed by other authors, including Laherrère (1999, 2003, 2004a, b, 2005). These techniques have much in common with those described above and raise a comparable set of issues and concerns. In principle, the extrapolation of discovery trends should provide more reliable estimates of the URR because the discovery cycle is more advanced (Fig. 5). However, discovery data is less accessible and reliable than production data and, unlike the latter, is estimated to different levels of confidence. Of particular importance is (a) whether the discovery estimates are based on *proved reserves* or *proved and probable reserves*; and (b) how revisions to
reserve estimates are reflected in the discovery data. These points are reviewed in Box 2 and Box 3.

Hubbert’s discovery projections were based on the idealised life-cycle model illustrated in Figure 5 (Hubbert, 1959). Hubbert assumed that both cumulative discovery and cumulative production grew logistically, with the former preceding the latter by some time interval. As the peak rate of discovery precedes the peak in production, identification of the former could form a basis for predicting the latter.

Hubbert (1962) fitted a logistic curve to the US data on cumulative proved discoveries and used this to estimate an URR of 170 Gb. Subsequent studies confirmed this estimate (Hubbert, 1966, 1968, 1974, 1979, 1982) but several authors questioned Hubbert’s results. For example, Ryan (1965, 1966) fitted logistic curves to the US production and discovery data and found they led to widely different estimates for the URR. He also showed that much larger estimates could be cited with equal justification and that the estimates increased rapidly with the addition of only a few more years of data. Cavallo (2004) recreated Hubbert’s original dataset and found that reserve growth at existing fields (Box 2). In order to take account of this, Laherrère (1996, 2002a, b) and others use backdated estimates which are typically larger than those made at the time of field discovery as a result of reserve growth in the intervening period (Box 2). Hubbert (1967) was one of the first persons to develop backdated discovery estimates, but did not use these for discovery projection. A discovery cycle based on backdated estimates will be a different shape from one based on current estimates and will have a different date for the peak in discoveries.

Backdated estimates provide a more accurate picture of what was found at a particular time and are also more suitable for estimating the URR since the cumulative discovery curve is more likely to trend to an asymptote. However, they can be misleading since the sizes of fields discovered at

The assumption that the discovery cycle takes the same form as the production cycle appears neither necessary nor plausible—although it works fairly well for the US when proven reserve data are used. The factors influencing discovery at different points in time are likely to be different from those influencing production at a later point in time and the skewed field-size distribution would be expected to (and frequently does) lead to a sharply rising cumulative discovery cycle and an asymmetric discovery cycle (Nehring, 2006a, b, c). For similar reasons, there is unlikely to be a predictable time lag between the peaks in discovery and production.

While Hubbert used proven (1P) reserves to form his cumulative discovery estimates, subsequent authors use proven and probable (2P) reserves (Campbell, 1997; Laherrère, 2004a, b; Bentley and others, 2007). Since these are generally larger than 1P estimates, they should lead to a higher estimate of the URR. Also, cumulative discovery estimates tend to increase over time even if no new fields are found as a result of reserve growth at existing fields (Box 2). In order to take account of this, Laherrère (1996, 2002a, b) and others use backdated estimates which are typically larger than those made at the time of field discovery as a result of reserve growth in the intervening period (Box 2). Hubbert (1967) was one of the first persons to develop backdated discovery estimates, but did not use these for discovery projection. A discovery cycle based on backdated estimates will be a different shape from one based on current estimates and will have a different date for the peak in discoveries.

Backdated estimates provide a more accurate picture of what was found at a particular time and are also more suitable for estimating the URR since the cumulative discovery curve is more likely to trend to an asymptote. However, they can be misleading since the sizes of fields discovered at

12...Because petroleum exploration in the US began very early, because the initial exploration and discoveries occurred in what has proved to be relatively minor basins, because early drilling technology was very limited in its drilling depth capabilities, and because discoveries in the major basins only hit their stride between 1910 in 1950, the US comes closest to a symmetric discovery curve of any major oil producing country or region’ (Nehring, 2006a, b, c).
Box 2. Understanding reserve estimates

Oil reserves are those quantities of oil in known fields which are considered to be technically possible and economically feasible to extract, under defined conditions. Reserve estimates are inherently uncertain and are quoted to three levels of confidence, namely proved reserves (1P), proved and probable reserves (2P) and proved, probable and possible reserves (3P). However, these terms are defined and interpreted in different ways and regulatory bodies have made only limited progress towards standardisation (Thompson and others, 2009a, b). Publicly available data sources are poorly suited to studying oil depletion and their limitations are insufficiently appreciated (Bentley and others, 2007). The databases available from commercial sources are better in this regard, but are also expensive, confidential and not necessarily reliable for all regions. Only a subset of global reserves is subject to formal reporting requirements and this is largely confined to the reporting of highly conservative 1P data for aggregate regions. Country-level 2P reserve estimates are available from commercial sources but only 1P data is available for the US. In the absence of audited estimates for individual fields, analysts must rely on assumptions whose level of confidence is inversely proportional to their importance—being lowest for those countries that hold the majority of the world’s reserves.

The common practice of adding 1P estimates to form a regional or global total is statistically incorrect and likely to significantly underestimate actual 1P reserves. Aggregation of 2P estimates should introduce less error but this may be either positive, negative or zero depending on the interpretation of the estimates and the shape of the underlying probability distributions—which is rarely available. Reserve estimates of known fields commonly undergo repeated revisions as a result of better geological understanding, improved extraction technology, variations in economic conditions and changes in reporting practices. Changes in reserve estimates over time are usually referred to as reserve additions, although the changes could be either positive or negative.

Box 3. Understanding cumulative discovery estimates

The cumulative discoveries in a region in a particular point in time may be estimated from the sum of cumulative production and declared reserves. Depending upon the data available, it may be possible to estimate cumulative 1P, 2P or 3P discoveries. The rate of discovery, or more simply discovery, is the first derivative of cumulative discovery with respect to time.

Cumulative discoveries are not changed by production since this merely transfers resources from one category (reserves) to another (cumulative production). However, cumulative discoveries may be either increased or reduced by revisions to the reserve estimates for known fields. The latter is usually referred to as reserve growth since estimates are normally revised upwards rather than downwards (Thompson and others, 2009a, b). However, a more accurate term is cumulative discovery growth, since reserves are continually being depleted by production. An alternative terms is ‘ultimate recovery growth’ since what is growing are the estimates of what will ultimately be recovered from the field or region. A growth function is a plot of the growth of cumulative discovery estimates for a particular category of field over time (Verma, 2003; Klett and Gautier, 2005; Verma, 2005). In some cases, the growth can be substantial, especially if 1P data is used. For example, Lore and others (1996) found that the estimated size of offshore fields in the Gulf of Mexico doubled within six years of discovery and quadrupled within 40 years. In principle, the reserve growth observed using 2P estimates should be less than this - although analysis of country-level data shows that this is not necessarily the case (Sorrell and others, 2009).

Some data sources record reserve revisions in the year in which they are made and make no adjustment to the discovery data for earlier years. Others backdate the revisions to the year in which the relevant fields were discovered. The logic of the first approach is that the reserves did not become ‘available’ for production until the estimate was revised and therefore should only appear at the time of the revision. The logic of the second approach is that the reserves are contained in a field that was discovered many years earlier, so backdating provides a more accurate indication of what was ‘actually’ found at that time as well as what will ultimately be recovered from that field. Both of these approaches have their merits, but the difference between them is not always appreciated. When cumulative discovery estimates are backdated, the ‘shape’ and ‘height’ of the cumulative discovery curve will change as revisions are made (Fig. 6).

different times will not have been estimated on a consistent basis (i.e. they will reflect differing amounts of reserve growth). For the same reason, both the height and shape of the curve will change over time, and they will not represent the ultimate resources that were found since more reserve growth can be anticipated in the future (Fig. 6). Hence, to provide reliable estimates of the URR, backdated discovery data should be adjusted to allow for future reserve growth (Drew and Schuenemeyer, 1992; Root and Mast, 1993). While there will be uncertainty regarding the appropriate growth function to use (Box 2), the failure to do so is likely to lead to underestimates of the URR. Lynch (2002), for example, likens the neglect of future reserve growth to: “...comparing old orchards with newly planted saplings and extrapolating to demonstrate declining tree size.” Campbell and Laherrère claim that adjustments are unnecessary, since 2P estimates should not change much following field discovery. However, this is inconsistent with the available evidence (Sorrell and others, 2009) and also contradictory, since, if 2P estimates are relatively stable, then there should be no advantage in backdating.
The complications introduced by reserve growth are illustrated by Nehring’s study of the Permian Basin and San Joaquin valley in the US—which both have been producing oil for more than 80 years (Nehring, 2006a, b, c). Nehring employs backdated cumulative proved discovery estimates and corrects these with Hubbert’s (1967) growth function to estimate the ultimate resources discovered in each time interval. When using data through to 1964, the corrected cumulative discovery curve for the Permian Basin suggests an URR of 27.5 Gb, compared to only 19 Gb with the uncorrected data. However, when using data through to 2000, the URR estimate is 37% larger, and the estimated date of peak discovery has moved back in time. While Hubbert’s growth function predicts substantial reserve growth, it nevertheless underestimates the growth that actually occurred—especially for the older fields. Nehring comments:

"...the continuous upward movement in the [corrected] cumulative discovery curve makes this curve useless as a tool for predicting the ultimate recovery. Estimates of ultimate recovery derived from cumulative discovery curves are only valid if one can guarantee that there will be no further increases in the ultimate recovery of discovered fields...no such guarantee can be made." (Nehring, 2006a, b, c)

This example could be unrepresentative, however. Proven reserve estimates would be expected to grow by more than 2P estimates since they represent a more conservative estimate of recoverable reserves. As a result, the estimates derived from 2P data should be more reliable. Also, Nehring relies on a growth function that is nearly 40 years old and is only applied to the most recent 30 years of data—despite more recent growth functions being available (Verma, 2003, 2005). In addition, the observed reserve growth derives primarily from CO2 injection in large fields of low permeability and such techniques may neither be suitable nor available in other fields and regions. However, Nehring highlights an important problem that is common to all techniques that rely on discovery data.

**DISCOVERY OVER EFFORT TECHNIQUES**

If data are available, then exploratory effort should provide a better explanatory variable than time, since the corresponding rate of discovery should be less affected by time-varying factors such as changes in tax regimes. Hubbert (1967) was one of the first persons to fit curves to discovery data as a function of exploratory effort and variants of this approach have subsequently been employed by other authors (Cleveland, 1992; Campbell, 1996; Ivanhoe, 1996; Laherrère, 2002a, b). This approach is also the basis of *discovery process modelling*, which was first introduced by Arps and Roberts (1958) and has subsequently been widely used (Power and Fuller, 1992a, b; Drew and Schuenemeyer, 1993; Drew, 1997). Both methods rely on backdated discovery estimates and require some method for estimating future reserve growth. However, only the latter needs data on individual fields.

There are several ways of measuring exploratory effort,13 although the choice will be largely dictated by data availability. The most common metric (the cumulative number of 'new field wildcat' wells (NFWs)) may not be the best, however, since much reserve growth derives from development rather than exploratory drilling (Cleveland, 1992). There are also difficulties when accounting for the delays between drilling and reserve additions, in distinguishing between the search for oil and the search for gas resources and in allowing for spatial and temporal variations in drilling patterns (Byrd and others, 1985).

---

13Including the cumulative length of exploratory drilling (Hubbert, 1967), the total number of exploratory wells (Ryan, 1973), the number of successful exploratory wells (Moore, 1962), the cumulative length of successful exploratory wells (Stitt, 1982) or the cumulative length of all wells (i.e. both exploratory and development) (Cleveland, 1992). A distinction may also be made between the first exploratory well to be drilled ('new field wildcats' or NFWs) and subsequent wells.
A ‘creaming curve’ is a plot of backdated cumulative discoveries against exploratory effort (Fig. 7), while a ‘yield per effort’ (YPE) curve is a plot of the rate of discovery against exploratory effort (i.e. the first derivative of the creaming curve) (Fig. 8). Provided the ‘yield’ from drilling declines as exploration proceeds, an estimate of the URR may be derived from the asymptote of the former, the integral of the latter or the corresponding parameters in the fitted curves. Changes in yield represent the net effect of changes in the success rate (the fraction of exploratory wells drilled that yield commercially viable quantities of oil) and changes in the average size of discovered fields. Evidence suggests that the success rate in most regions has declined only relatively gently, if at all, indicating that improvements in exploration technology have partially or wholly offset the anticipated decline in the success rate as a result of the declining number of undiscovered fields (Meisner and Demirmen, 1981; Forbes and Zampelli, 2000). However, since large fields generally occupy a larger surface area, they tend to be found relatively early even if drilling is random (Arps and Roberts, 1958). In contrast, the average size of discovered fields in many regions has fallen by an order of magnitude since the early days of exploration. Hence, declining YPE is most likely to be the result of falling average field-sizes.

Hubbert’s investigation of YPE curves was developed in response to Zapp (1962), who used exploratory effort as an explanatory variable but assumed that the yield would remain unchanged, leading to an unrealistically large estimate for the US URR (590 Gb). Zapp’s approach was subsequently adopted by Hendricks (1965), who simply assumed that the yield would decline linearly. In contrast, Hubbert (1967) based his forecast of future yield on a detailed analysis of past trends, which showed a negative exponential decline.

Hubbert (1967) fitted a negative exponential curve to his estimates of YPE in the lower 48 US states and estimated a URR of ~170 Gb, consistent with his estimates from production and discovery projection. However, Harris (1977) showed that Hubbert’s method violated standard statistical procedures, placed excessive weight on the last (and most uncertain) data point and led to systematically biased estimates. The only reason Hubbert’s estimate was consistent with his earlier study was that the discovery rate had increased—something which Hubbert considered to be both anomalous and temporary. Harris also showed that a YPE curve for an aggregate region such as the US will not necessarily be exponential, even if the trends for individual regions are exponential. As a result, the URR estimated from a curve fit to aggregate data will be different from the sum of the estimates from curves fit to regional data.

A few recent studies have used creaming curves rather than YPE curves. Laherrère (2001, 2002a, b, 2004a, b) has estimated creaming curves for all the regions of the world and found that they tend to rise steeply in the early stages of exploration, reflecting the discovery of small number of large fields. He fits ‘hyperbolas’ to these data, but rarely provides either the functional form or the goodness of fit. Smooth curves (e.g. Fig. 1) may be the exception rather than the rule, however. For example, Sneddon and others (2003) show how the YPE in an exploration play typically exhibits two or three plateaus and provide some technical reasons as to why this may be the case. While diminishing

![Cumulative Discovery (Gb)](image_url)

Figure 7. Example of a ‘creaming’ curve. Source: IHS Energy. Note: Name of region withheld on grounds of data confidentiality.

---

[14] Laherrère (2004a, b) states that the creaming curve was invented by Shell in the 1980s, but variants of this approach have been used in the oil industry for much longer period (Arps and Roberts, 1958; Arps and others, 1971; Odell and Rosing, 1980; Harbaugh and others, 1995). Two employees of Shell published a paper on ‘the creaming method’ in 1981, but this describes a highly sophisticated (and not widely used) discovery process model that relies on Monte Carlo simulation of trends in both success rates and average field sizes and assumes a lognormal field size distribution (Meisner and Demirmen, 1981).

returns to exploratory effort are widely observed at the play level, the same may not always be observed at larger geographical scales because regions are frequently developed in the order of ease of exploration and development rather than size (Charpentier, 2003). For example, a combination of geological accessibility and improvements in exploration technology led to the largest play in the Michigan basin being developed relatively late (Charpentier, 2003). Similar phenomena are reported by Wendebourg and Lamiraux (2002), who find two exploration cycles in the Paris basin. While a creaming curve estimated using data through to 1986 leads to an URR estimate of 15Mt, a similar curve estimated using data through to 1996 leads to a much larger estimate of 46Mt. The larger the geographical region, the more significant this problem could become. Laherrère (2003, 2004a, b) addresses this through the use of multi-cycle models, but provides little statistical support for his choice of curves and, in some cases, the appropriate number is unclear. For example, Laherrère (2003) models the oil and gas resources of the Middle East with two creaming curves but in a subsequent article, this has increased to four (Laherrère, 2004a, b). The choice appears to be largely determined by the ‘shape’ of the data, with interpretation in terms of exploration history being made ex post.

The potential for further exploration cycles cannot be established from the statistical analysis of historical data, but only from a detailed evaluation of geological potential and exploration history. For small, geologically defined regions where exploration is well advanced, the probability of new cycles may be relatively low, while for large, politically defined regions, which are partly unexplored (e.g. owing to the depth of drilling required, or geographical remoteness, or political restrictions) the probability may be much higher. The reliability of curve-fitting, therefore, depends heavily on the assumption that any new exploration cycles will have only a small impact on aggregate resources—either because there will be no or few such cycles, or because the discovered resources will be relatively small. While this judgement may be reasonable for many regions around the world, it remains problematic for key regions such as Iraq. Unfortunately, these are precisely the regions that account for a significant proportion of the global URR.

In sum, while exploratory effort should provide a better explanatory variable than time, curve-fitting techniques must still be used with care. Common difficulties include the inaccuracy of the data used for discovery estimates, the uncertainty about future reserve growth, the apparent sensitivity of the results to the choice of functional form, the existence of multiple exploration cycles and the inability to anticipate new exploration cycles in the future. Overall, these difficulties appear more likely to lead to underestimates of the regional URR.
CONSISTENCY OF CURVE-FITTING TECHNIQUES

We investigated the reliability of curve-fitting techniques with the help of illustrative data from 10 oil-producing regions. The data were taken from a database supplied by IHS Energy and includes annual estimates of production, backdated 2P discoveries and the number of new field wildcat wells for all oil-producing countries. The selected regions include both individual countries and groups of countries, with all but one (Region B) apparently past their peak of discovery, and five past their peak of production. Since the objective was to test the reliability of curve-fitting techniques as currently used, we did not correct the discovery estimates to allow for future reserve growth.

We estimated the URR for each region using production decline curves, cumulative discovery projections and creaming curves. In each case, we investigated the consistency of the estimates obtained with different lengths of data series, different choices of functional form and different numbers of curves, and also compared the estimates produced by each technique. For illustrative purposes, we judged two sets of results to be consistent if the mean URR estimates differed by less than 20% of the cumulative production \((Q_{2007})\) or cumulative discoveries \((D_{2007})\) in the region through to 2007. A more or less stringent definition of consistency would not significantly change the results, since most estimates were found to be either broadly consistent or substantially different. The full results of these tests are given in Sorrell and Speirs (2009) and are summarised below.

First, the results raise concerns about the reliability of the URR estimates from curve-fitting techniques, at least when (as is usually the case) they are applied at the country or regional level with data that has not been corrected for future reserve growth. In particular, we observed that: (a) in only one of the regions examined were the mean URR estimates consistent between all three curve-fitting techniques; (b) variations in the length of time series, functional forms and number of curves led to inconsistent results more often than consistent results; and c) the degree of inconsistency in the URR estimates was frequently very large. While estimates were more likely to be consistent for regions at a later stage of their discovery and/or production cycle, inconsistent results were frequently obtained for mature regions as well.

Different functional forms were often found to fit the data equally well, but to provide substantially different estimates of URR. This is illustrated in Figure 9 which shows cumulative discovery projections for each region using both logistic and Gompertz functional forms. Taking Region E as an illustration, the difference between the \(R^2\) for each model is only 0.003 but the URR estimates differ by a third. The mean difference in the URR estimates from the two models was 59% of the cumulative discoveries through to 2007 (ranging from 1% to 362%), but the mean difference in \(R^2\) estimates was only 0.001. The estimates only converge when regions are at a relatively late stage in their discovery cycle when the asymptote of the curve is clearly apparent (e.g. Region I). As a result, estimates made at earlier stages in the discovery cycle can lead to significantly different results (see Fig. 10). Contrary to expectations, a logistic or Gompertz functional form was found to be more appropriate than an exponential in all the regions examined, but the choice can bias the results (e.g. the Gompertz model provides higher estimates in all cases).

The production decline (‘Hubbert Linearisation’) technique was found to be particularly unreliable and the results suggest a systematic tendency to underestimate the URR. As an illustration, Figure 11 shows the results for Region J. While the onshore data ‘settles’ into an approximate linear relationship that can be modelled by a single linear regression, the corresponding data for both offshore and the region as a whole exhibits ‘trend breaks’ that may result from additional cycles of exploration and production. If this technique had been used at an earlier stage of the production cycle (i.e. prior to the trend break), then it would have led to a significant underestimate of the regional URR. Trend breaks of this form were observed for six of the ten regions using aggregate data and for all of the regions using either onshore or offshore data (or both). The frequency of such breaks gives little confidence that the decline curves will remain stable in the future.

\[\text{Consistency of curve-fitting techniques}\]

\[\text{The names of the regions are withheld due to data confidentiality.}\]
Figure 9. Cumulative discovery projection results.
Our results also do not support the claim that creaming curves are generally more reliable than discovery projections. Notably, the creaming curves for four of the regions do not exhibit asymptotic behaviour, although, in two of these cases, the corresponding discovery projection was asymptotic. Once again, the results were sensitive to the particular functional form assumed (see Fig. 12) and while three of the regions could be fit with either one or two creaming curves, the corresponding URR estimates were significantly different (see Fig. 13). Without a detailed knowledge of the exploration history of a region, it is difficult to justify one choice over the other.

The primary reason for these inconsistent results is that the techniques are being applied to large and geologically diverse regions that lack a consistent exploration history. In addition, we did not always distinguish between onshore and offshore regions, the data source did not classify exploratory drilling as searching for either oil or gas, and we did not correct the discovery data to allow for future reserve growth. Future applications of curve-fitting should, therefore, address each of these problems as far as the available data permits. Nevertheless, the results are sufficient to demonstrate the limitations of curve-fitting technique as currently used and suggest that the associated URR estimates should be treated with caution.

**RECONCILING CURVE-FITTING WITH ECONOMETRICS**

Curve-fitting techniques assume that the ‘shape’ of the production or discovery cycle can be estimated from the historical data and that this shape will not be significantly affected by any future changes in prices, technology and other relevant variables. As a result, there has been a tendency to neglect these variables, despite the potential errors that may result. For example, low oil prices and political constraints may restrict production when resources are abundant, while high prices may stabilise or increase production when depletion is advanced. As a result, many applications of curve-fitting techniques are likely to suffer from missing variable bias and/or serial correlation of the error terms. This could lead to biased estimates of model parameters.
parameters (including the URR), underestimates of the associated standard errors and overestimates of the model goodness of fit. Several examples of this are provided in Sorrell and Speirs (2009).

These problems may potentially be addressed by including one or more ‘lags’ of the dependent variable within the model specification (e.g. making production in the current year a function of production in one or more previous years), but the re-specified model may not necessarily lend itself to the estimation of URR. A more promising approach is to include some of the economic and political determinants of discovery and/or production within the model specification. This effectively gives a ‘hybrid’ of a curve-fitting approach to estimating URR and an econometric approach to estimating future oil discoveries (Walls, 1992, 1994). The latter originated with Fisher (1964), who estimated equations for exploratory activity, success rate and the average size of discovered fields as a function of oil prices, past average discovery size and lagged dependent variables. Many variants have followed, but these have mostly focused on forecasting discoveries rather than estimating URR and are also likely to be biased since they ignore the geological determinants of discoveries and production (Power and Fuller, 1992a, b).

A good example of a hybrid model is Kaufmann (1991), who fits a logistic model to the US cumulative production and then uses an econometric model to account for the deviations between predicted and actual production. This formulation assumes that geologic and physical factors cause oil production to rise and fall over the long term, while economic and political variables (e.g. legal restrictions on production)19 lead to short-term variations around this underlying trend. This model provides a much better fit to US production over the period 1947 to 1985, but estimates of the URR require assumptions about the future values of the relevant economic and political variables. Pesaran and Samiei (1995) argue that Kaufmann’s model is biased because the estimation of URR in the first stage does not take into account the effect of economic factors which only enter the analysis in the second stage. They avoid this by modifying the logistic model to allow for the dependence of URR on a number of economic and other variables and re-formulating it to eliminate problems of serial correlation. Their model explains over 98% of the variation in the US production over the period of 1948–1990 and leads to a higher estimate of the URR. However, the oil price elasticity of the URR is both symmetric and independent of time (e.g. the same before and after the peak) which seems implausible.

---

18For example, Epple and Hansen (1981), MacAvoy and Pindyck (1973), Walls (1994) and Mohn and Osmundsen (2008).

19For example, between 1957 and 1968, the ‘prorationing’ decisions of the Texas Railroad Commission shut in more than 50% of Texan oil-producing capacity.
In a significant innovation, Kaufmann and Cleveland (2001) include the average US production costs as an explanatory variable and hence remove the need to assume a particular functional form for the production cycle. Their empirically estimated U-shaped average cost curve mirrors the bell-shaped production curve and represents the net effect of resource depletion and technical change— with steeply rising costs after 1970 indicating accelerating depletion (Cleveland, 1991). This model accounts for most of the variation in the US oil production between 1938 in 1991 and highlights the importance of economic and political variables in determining the 'shape' of the discovery or production cycle:

"...Hubbert was able to predict a peak in US production accurately because real oil prices, average real cost of production, and decisions by the Texas Railroad Commission coevolved in a way that traced what appears to be a symmetric bell-shaped curve for production over time. A different evolutionary path for any of these variables could have produced a pattern of production that was significantly different from a bell shaped curve... In effect, Hubbert got lucky." (Kaufmann and Cleveland, 2001)

However, while this model overcomes a key weakness of curve-fitting techniques (i.e. the arbitrary choice of functional form), it may not be applicable for other regions owing to the lack of data on production costs. Moreover, if it is to be used to estimate the regional URR, then some assumption is required about future trends in production costs. In addition, while Hubbert may have 'got lucky' in forecasting the date of peak production, it is much less clear as to whether accounting for economic variables will make a significant difference to the estimated URR.

If the required data are available, then very similar approaches can be used to modify and improve any of the curve-fitting techniques. For example, Cleveland and Kaufmann (1991, 1997) modified Hubbert's exponential model of YPE to account for short-term changes in oil prices and the rate of drilling. Their equation provides a much better fit to historical trends and shows how periods of relative stability or increases in YPE were associated with changes in the rates of drilling and/or oil prices. However, again, depletion dominates in the long-term, implying that the revised model may not significantly change the estimates of URR (Kaufmann and Cleveland, 1991).

Some authors attempt to separate the effect of technical change from that of resource depletion (Power and Jewkes, 1992; Iledare and Pulsipher, 1999). For example, in their study of YPE in the Gulf of Mexico, Managi and others (2005) model depletion by cumulative discoveries and technical change by an index of the annual number of innovations adopted by the offshore industry, weighted by their relative importance (NPC, 1995; Managi and others, 2004). The results show that the pace of technical change has increased since 1975, greatly expanding the area of exploration and leading to a YPE in 2000 that is comparable to that achieved 50 years previously. While the geological diversity of the Gulf of Mexico contributes to this result, a more likely reason is that exploration has been geographically restricted in the past, owing to a combination of the technical difficulties of deep-water drilling and changing licensing regimes (Priest, 2007). As a result, fields have not been found in the approximate declining order of size that is normally assumed.

In sum, while hybrid models improve on standard curve-fitting, they may be more suitable for short-term supply forecasting than for estimating URR. The latter requires assumptions about the future values of the relevant explanatory variables and despite their better fit to historical data, it is not obvious that hybrid models lead to substantially different estimates of URR. However, they do allow the dependence of the URR on energy prices and other factors to be directly explored.

Hybrid models may still lead to misleading conclusions if applied to regions that lack either geological homogeneity or a consistent exploration history and may still be vulnerable to missing variable bias since it is impractical to include more than a subset of the variables that could affect production and/or discovery trends (e.g. tax rules, leasing decisions, geographical restrictions on exploration, production/import/export quotas etc.) (Lynch, 2002). Also the required data are frequently either lacking or unreliable and are rarely available at the level of disaggregation required. Problems such as these may partly explain why there are so few 'hybrid' studies, and why the available studies are largely confined to the United States.20

**SUMMARY**

M. King Hubbert has left an important and influential legacy in many areas, including in

---

20 A notable exception is Dées and others (2007).
particular, the use of curve-fitting methods to estimate URR. These techniques were central to the 1960s debate on the US oil production and remain equally relevant to the contemporary debate on 'peak oil'. This article has reviewed these techniques and compared them with other approaches to estimating URR. The main conclusions are as follows.

First, curve-fitting forms part of a broader group of techniques based on the extrapolation of historical trends in production and discovery. These techniques differ in degree rather than kind and share many of the same strengths and weaknesses. However, a key practical difference is that field-size distribution and discovery process techniques require data on individual fields, while simple curve-fitting only requires aggregate data at the regional level. All these techniques assume a skewed field size distribution and diminishing returns to exploration, with the large fields being found relatively early. However, these assumptions will only hold if depletion outweighs the effect of technical change and if the region is geologically homogeneous and has had a relatively uninterrupted exploration history. Unfortunately, this is rarely the case.

Second, many applications of curve-fitting take insufficient account of the weaknesses of this technique, including: the inadequate theoretical basis; the sensitivity of the estimates to the choice of functional form; the risk of overfitting multi-cycle models; the inability to anticipate future cycles of production or discovery; and the neglect of economic, political and other variables. In general, these weaknesses appear more likely to lead to underestimates of the URR and have probably contributed to excessively pessimistic forecasts of oil supply. Curve-fitting to discovery data introduces additional complications such as the uncertainty in reserve estimates and the need to adjust estimates to allow for future reserve growth. The common failure to make such adjustments is likely to have further contributed to underestimates of resource size.

Third, tests of curve-fitting techniques using illustrative data from a number of regions have shown how different techniques, functional forms, length of time series and numbers of curves can lead to inconsistent results. However, while this raises concerns about the reliability of such estimates, the degree of uncertainty declines as exploration matures. Hence, as more and more regions become extensively explored, the accuracy of regional and global URR estimates should improve. In addition, some of the limitations of curve-fitting may be overcome with the use of hybrid models that incorporate relevant economic and political variables. However, despite their better fit to historical data, such models may not lead to substantially different estimates of the URR.

Finally, while these limitations do not mean that curve-fitting should be abandoned, they do imply that the applicability of these techniques is more limited than some adherents claim and that the confidence bounds on the results are wider than is usually assumed. Where possible, resource assessments should employ multiple techniques and sources of data and be informed by knowledge of the geological characteristics and exploration history of the region. They must also acknowledge the considerable uncertainty in the results obtained.

ACKNOWLEDGMENTS

This article is based on a comprehensive review of the curve-fitting literature by Sorrell and Speirs (2009) which in turn formed part of wide-ranging study of global oil depletion by the UK Energy Research Centre (Sorrell and others, 2009). The authors would like to thank the UK Research Councils for their financial support, IHS Energy for allowing the publication of data from their PEPS database, Fabiana Gordon for her help on statistical issues, and Roger Bentley, Richard Miller, Jean Laherrère and Robert Kaufmann for their helpful comments on earlier drafts. The usual disclaimers apply.

REFERENCES


Cramer Barton, C., and La Pointe, P. R., 1995. Fractals in petroleum geology and earth processes: Springer.


Hubbert, M. K., 1956. Nuclear energy and the fossil fuels, in Meeting of the Southern District, Division of Production, American Petroleum Institute, San Antonio, TX: Shell Development Company.


Shaping the global oil peak: A review of the evidence on field sizes, reserve growth, decline rates and depletion rates

Steve Sorrell a,*, Jamie Speirs b, Roger Bentley c, Richard Miller d, Erica Thompson b

a Sussex Energy Group, SPRU (Science & Technology Policy Research), Freeman Centre, University of Sussex, Falmer, Brighton BN1 9QE, UK
b Imperial College Centre for Energy Policy and Technology, London, UK
c Department of Cybernetics, University of Reading, UK
d Oil Depletion Analysis Centre, London, UK

A R T I C L E   I N F O

Article history:
Received 13 July 2011
Received in revised form 7 October 2011
Accepted 9 October 2011
Available online xxx

Keywords:
Oil depletion
Peak oil
Reserve Growth
Decline rates
Depletion rates

A B S T R A C T

This review paper summarises and evaluates the evidence regarding four issues that are considered to be of critical importance for future global oil supply. These are: a) how regional and global oil resources are distributed between different sizes of field; b) why estimates of the recoverable resources from individual fields tend to grow over time and the current and likely future contribution of this to global reserve additions; c) how rapidly the production from different categories of field is declining and how this may be expected to change in the future; and d) how rapidly the remaining recoverable resources in a field or region can be produced. It is shown that, despite serious data limitations, the level of knowledge of each of these issues has improved considerably over the past decade. While the evidence on reserve growth appears relatively encouraging for future global oil supply, that on decline and depletion rates does not. Projections of future global oil supply that use assumptions inconsistent with this evidence base are likely to be in error.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Conventional oil (namely crude oil, condensate and natural gas liquids) currently accounts for over 97% of global liquid fuels production and is still expected to account for around 90% in 2030 [1]. But many commentators are forecasting a near-term peak and subsequent terminal decline in the production of conventional oil, with alternative sources being unable to ‘fill the gap’ on the timescale required [2,3]. In contrast, others argue that liquid fuels production will be sufficient to meet global demand well into the 21st century, as rising prices stimulate new discoveries, enhanced recovery and the development of non-conventional resources such as oil sands [4–6]. The topic of ‘peak oil’ is notoriously contentious and confused and both the academic and policy debates remain polarised. However, knowledge is growing in key areas [7] and many companies [8], governments [9–11], commentators [12,13] and international organisations [14] are becoming increasing pessimistic about future supply. For example, the IEA (International Energy Agency) has steadily reduced its long-term projections for global oil supply and now acknowledges that the global production of crude oil1 has past its peak [15–17].

Oil discovery and production is shaped by multiple geological, technical, economic and political factors that combine to create considerable uncertainty over future supply. Disagreement over medium to long-term supply projections is further influenced by competing disciplinary orientations, methodological disputes, inadequate data and conflicting evidence over the contribution of key variables such as ‘reserve growth’ [7]. Nevertheless, there is potential for increasing the degree of consensus in a number of areas and considerable progress has been made over the last few years. The objective of this paper is to summarise the current state of knowledge regarding four issues which are expected to have a critical influence on the timing and shape of the global oil peak, namely:

• Field-size distributions: how regional and global oil resources are distributed between different sizes of field and the relative importance of large and small fields.
2. Field-size distributions

Methods of resource assessment and supply forecasting frequently rely upon assumptions about the size distribution of oil fields within a region, where ‘size’ refers to the estimated URR (ultimately recoverable resources) of each field. It is well established that:

- the majority of recoverable resources within a region tend to be contained within a small number of large fields; and
- these large fields tend to be discovered relatively early, with subsequent discoveries being progressively smaller and requiring more effort to locate.

This rule seems broadly applicable at levels ranging from individual ‘exploration plays’\(^2\) to the entire world. However, the precise form of the size distribution varies from one region to another and is a long-standing focus of dispute [21–23]. Of particular interest is the proportion of resources contained within very large and very small fields, since the former dominate current oil production and the latter are expected to become increasingly important in the future.

### 2.1. The global importance of large fields

One of the first global surveys of crude oil fields was by Ivanhoe and Leckie [24] who grouped fields into ten size categories on the basis of their estimated URR (Table 1). The 370 fields with a URR exceeding 0.5 Gb (billion barrels) represented less than 1% of the total number of fields but accounted for three quarters of all the oil that had ever been discovered (cumulative discoveries). Global production of crude oil was approximately 70 million barrels of oil per day (mb/d) in 2008, so 0.5 Gb corresponds to approximately 7 days of global supply. Of particular importance were the 42 ‘super-

<table>
<thead>
<tr>
<th>Category</th>
<th>Estimated URR (mb)</th>
<th>No. in world</th>
</tr>
</thead>
<tbody>
<tr>
<td>Megagiant</td>
<td>&gt;50,000</td>
<td>2</td>
</tr>
<tr>
<td>Super-giant</td>
<td>5000–50,000</td>
<td>40</td>
</tr>
<tr>
<td>Giant</td>
<td>500–5000</td>
<td>328</td>
</tr>
<tr>
<td>Major</td>
<td>100–500</td>
<td>961</td>
</tr>
<tr>
<td>Large</td>
<td>50–100</td>
<td>895</td>
</tr>
<tr>
<td>Medium</td>
<td>25–50</td>
<td>1109</td>
</tr>
<tr>
<td>Small</td>
<td>10–25</td>
<td>2128</td>
</tr>
<tr>
<td>Very small</td>
<td>1–10</td>
<td>7112</td>
</tr>
<tr>
<td>Tiny</td>
<td>0.1–1</td>
<td>10,849</td>
</tr>
<tr>
<td>Insignificant</td>
<td>&lt;0.1</td>
<td>17,740</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>41,164</td>
</tr>
</tbody>
</table>

Source: [24].

Table 1

Ivanhoe and Leckie’s estimates of the size distribution of the world’s oil fields.

\(^2\) A ‘play’ is an area for petroleum exploration that has common geological attributes and lies within some well-defined geographic boundary.

\(^3\) These figures demonstrate that much more exploration has taken place in the US compared to other regions of the world and suggests there could be unexploited potential outside of the US.

\(^4\) Höök et al. [27] estimate there are 20 giant fields under Simmons’ definition which are not giant fields under Ivanhoe’s definition.
old, many are well past their peak of production and most of the rest will begin to decline within the next decade or so. The remaining reserves at these fields, their future production profile and the potential for reserve growth are therefore of critical importance for future global supply.

### 2.2. The future contribution of small fields

The size distribution of the population of oil fields in a region can be inferred from the size distribution of the sample of discovered fields. Arps and Roberts [29] were among the first to observe that this typically took a lognormal form — in other words, the frequency distribution of the logarithm of discovered field sizes resembled a normal distribution (Fig. 2). This observation was subsequently supported by several studies, including McCrossan’s [30] analysis of reservoir sizes in Western Canada and studies of US data by Kaufman [31] and Drew and Griffiths [32].

The notion that oil fields followed a lognormal size distribution became conventional wisdom during the 1960s and 1970s and subsequently formed the basis of some highly sophisticated ‘discovery process models’ [33–35]. This was despite the difficulties in drawing inferences from a relatively small sample of fields that may be unrepresentative of the population as a whole [36,37]. In particular small fields are likely to be underrepresented in such samples because they are less economic to develop — so-called ‘economic truncation’ [29,38–40]. The lognormal distribution may therefore result from the undersampling of small fields (Fig. 3). As exploration proceeds, technology improves, oil prices rise and/or costs fall, these fields should become increasingly economic to find and develop, causing the modal size of the sample of discovered fields to fall. This process has been observed in many oil-producing regions, including the Permian basin in Texas [21,41]. Additional sampling bias may be introduced by the tendency to discover the larger fields first [42].

On the basis of these and similar observations, Drew and colleagues proposed that the population field-size distribution was more likely to take a ‘power-law’, or ‘Pareto’ form (Box 1). If this is the case, a plot of the number (N) of fields exceeding a particular size (V) on logarithmic scales should approximate a straight line (Fig. 4). In contrast, if the size distribution is lognormal, this plot would be curved. While curved plots are more commonly observed in practice, this is partly the result of biased sampling.

Barton and Scholz [43] show how the power-law model provides a good fit to data from six regions ranging from a single exploration play to the entire world. However, Laherrère [23,44] argues that “…natural data gives rise to curved, not linear plots.” and shows how a quadratic cumulative frequency distribution (a ‘parabolic fractal’) can often provide a better fit. The curvature of these plots typically reduces over time as more small fields are found.

The proportion of the URR contained in smaller, undiscovered fields may be estimated by fitting one of these functions to the size distribution of discovered fields and extrapolating to smaller field sizes. The proportion will be greater with a power-law distribution, smaller with a parabolic fractal and smaller still with a lognormal. For example, Barton and Scholz [43] fitted a power-law to data from six regions and estimated that undiscovered small fields contained...
Box 1. Power-law field-size distributions and Zipf’s law

Let \( N(V) \) represent the number of fields that exceed a particular size \( V \). A power-law field-size distribution is given by: 

\[ N(V) = AV^{-\alpha}, \]

where \( A \) is a scaling factor and \( \alpha \) defines the shape of the distribution. Hence, a plot of the natural log of \( N \) against the natural log of \( V \) should approximate a straight line with slope \(-\alpha\).

\[ \ln N(V) = \ln A - \alpha \ln V \]

(Fig. 4). Barton and Scholz [43] fitted this equation to data from six regions and found values of \( \alpha \) ranging from 0.8 to 1.0. If \( \alpha \leq 1.0 \), the volume of oil in each field size class will decrease as the size class itself decreases, but as \( \alpha \) tends towards 1.0, small fields will account for increasing proportion of the URR.

Power-law distributions are also termed ‘Pareto distributions’, after Vilfredo Pareto [96] who represented income distribution in a similar way. They belong to a family of distributions known as ‘probabilistic fractals’ which have their roots in the work of Mandelbrot [97]. Indeed, Mandelbrot [98] was the first to propose that oil resources could be modelled with a power-law and demonstrated this by an analysis of US oil fields. Davies and Chang [99] have criticised the power-law model, but it has been widely used to assess the petroleum resources of the US [100,101].

Power-law distributions are related to ‘Zipf’s law’ which describes a relationship between the size and ‘rank’ \( N \) of discrete phenomena [102,103]. When oil fields are ranked in descending order of size so that the largest is rank 1, Zipf’s law states that the product of the rank and size is approximately constant \( (N(V))V \sim k \). The applicability of Zipf’s law is usually investigated by plotting cumulative frequency \( N(V) \) as a function of field size \( V \) [104]. The two approaches are equivalent.

3. Reserve growth

The phenomenon of reserve growth is critical to future global oil supply and an enduring topic of controversy within the ‘peak oil’ debate. To understand the issue, it is first necessary to review the definitions of reserves, reserve additions and cumulative discoveries.

Oil reserves are those quantities of oil in known fields which are considered to be technically possible and economically feasible to extract under defined conditions. Reserve estimates are inherently uncertain and are commonly quoted to two levels of confidence, namely proved reserves \( (1P) \) and proved and probable reserves \( (2P) \). These terms are defined and interpreted in different ways by different bodies, but they commonly imply a 90% and 50% probability respectively of recovered resources exceeding the stated figure [45]. Only a subset of global reserves is subject to formal reporting requirements and this is largely confined to the reporting of highly conservative 1P data for aggregate regions. Country-level 2P reserve estimates are available at cost from commercial sources but only 1P data is available for the US. Most reserve estimates are neither audited nor reliable, so analysts must rely upon assumptions whose level of confidence is inversely proportional to their importance – being lowest for those countries that hold the majority of the world’s reserves.6

Reserve estimates of known fields commonly undergo repeated revisions as a result of cumulative production, better geological understanding, improvements in extraction technology, variations in economic conditions and changes in reporting practices. Changes in reserve estimates over time are commonly referred to as reserve additions, although the changes could be either positive or negative.

The cumulative discoveries in a region in a particular point in time may be estimated from the sum of cumulative production and declared reserves. Cumulative discoveries are not changed by production since this merely transfers resources from one category (reserves) to another (cumulative production). However, cumulative discoveries may be either increased or reduced by revisions to the reserve estimates for known fields. This is commonly referred to as reserve growth since estimates are normally revised upwards rather than downwards [19]. However, a more accurate term is cumulative discovery growth, since reserves are continually being depleted by production. An alternative term is ‘ultimate recovery growth’ since what is growing are the estimates of what will ultimately be recovered from the field or region.

It is this process of reserve growth, rather than new discoveries that accounts for the majority of reserve additions in most regions of the world [19]. Most analysts expect this pattern to continue. However, reserve growth remains poorly understood.

---

Fig. 4. Log–log plot of cumulative frequency versus field size, illustrating the difference between theoretical population distribution and observed sample distribution.

---

6 The common practice of adding 1P estimates to form a regional or global total is statistically incorrect and likely to significantly underestimate actual 1P reserves [46]. Aggregation of 2P estimates should introduce less error but this may be either positive, negative or zero depending upon the statistical interpretation of the estimates and the shape of the underlying probability distributions – which is rarely available.

Please cite this article in press as: Sorrell S, et al., Shaping the global oil peak: A review of the evidence on field sizes, reserve growth, decline rates and depletion rates, Energy (2011), doi:10.1016/j.energy.2011.10.010
3.1. Sources of reserve growth

The multiple factors that contribute to reserve growth can be grouped under three headings, namely geological, technological and definitional.

Geological factors represent an increase (or decrease) in the estimates of OOIP (original oil in place) for a reservoir, field or region as a result of improved geological knowledge. For example, new reservoirs or extensions to previous reservoirs may be discovered or a better understanding of the volume, shape and characteristics of reservoirs may be obtained through the use of seismic and other techniques. In some cases, smaller fields that were previously classified as separate may be merged into larger fields as exploration proceeds [21]. This process may affect the historical data from which reserve growth is estimated, but it does not represent the growth of OOIP for an individual field.

Technological factors represent those activities which increase the estimated recovery factor, or the proportion of the OOIP that is both technically possible and economically viable to recover. Recovery factors can be as high as 80% for high permeability reservoirs, though the estimated global average is around 34% [1].

It is common to distinguish between primary recovery where oil is recovered under its own pressure, and secondary recovery where active pumping or water/gas injection is employed to increase reservoir pressures and the flow of oil to the surface. In some cases there is scope for substantially improving the recovery factor through the application of EOR (enhanced oil recovery) techniques whose suitability varies with the type, accessibility and characteristics of the field. Extensive use of EOR in the US has led to recovery factors that are ~5% higher than the global average and there may be considerable scope for application in other regions of the world.

In principle, each percentage increase in the recovery factor can be obtained through the application of EOR techniques. In some cases, smaller fields that were previously classified as separate may be merged into larger fields as exploration proceeds [21]. This process may affect the historical data from which reserve growth is estimated, but it does not represent the growth of OOIP for an individual field.

Definitional factors comprise a mix of definitional, legal, economic and political factors that influence reserve estimates but are independent of either the OOIP or recovery factor. These include changes in reserve classification schemes, such as occurred in Russia in the 1990s, and the practice of excluding reserves at discovered fields that have yet to receive production sanction. There may also be an implicit shift in definitions over time as a result of changes in personnel, operators and reporting cultures. All reserve definitions require assessment of economic viability, so changes in technology, oil prices and other economic conditions may also affect the volume of declared reserves.

The amount of reserve growth will depend upon the particular definition of reserves on which the cumulative discovery estimates are based (e.g. 1P or 2P). Reserve growth may be particularly high for cumulative discovery estimates based upon 1P reserves since these are a very conservative estimate of recoverable resources. Reserve growth should be smaller for cumulative discovery estimates based upon 2P reserves and if these correspond to median estimates of the URR, we would expect cumulative discovery estimates to be downgraded as frequently as they are upgraded. However, analysis suggests that this is not the case — estimates based upon 2P reserves normally increase over time [21].

Figs. 5 and 6 show the change in cumulative discovery estimates for fields in the UKCS (UK Continental Shelf). These estimates were published by the UK government and we take them to be similar to 2P. For large fields, the mean estimate of cumulative discoveries increased by approximately 50% over 27 years, while none of the individual estimates decreased in size (Fig. 5). If this is the case for 2P data from other regions of the world, we may expect significant reserve growth in the industry databases. However, smaller fields in the UKCS grew by only 20% over this period while many fields discovered since 1980 have shown a decrease (Fig. 6).

3.2. Estimating and forecasting reserve growth

Reserve growth can be estimated by comparing the initial booking of reserves for a specific field, with the sum of cumulative production and declared reserves for subsequent years. But the required data is only publicly available for a limited number of regions around the world. IHS Energy hold data on production and 2P reserves for most of the world’s fields, but this information is inaccessible to most analysts. Also, to estimate reserve growth it would be necessary to obtain this database for concurrent years and there are questions about its completeness prior to 2000.

Reserve growth has been most closely studied in the United States, where it accounted for 89% of the additions to US proved reserves over the period 1978 to 1990 [49–56]. While reserve growth also occurs in other regions of the world, the evidence base is thinner [56–68]. But despite being systematically investigated more than 40 years ago [69], reserve growth was relatively neglected before the 1980s [21]. An important stimulus to further investigation was the retrospective examination of discovery forecasts for the US, which were found to have systematically underestimated future discoveries as a result of neglecting reserve growth at known fields [75].

Future reserve growth can be estimated through the creation of reserve growth functions based upon the measured growth of a statistically significant sample of fields [54]. Both annual and cumulative growth functions can be calculated and used to convert current estimates of cumulative discoveries into future estimates for a specified year, with the amount of growth depending solely upon the age of the field. Fig. 7 shows two growth functions estimated for onshore US oil fields [50,51]. Both show rapid growth in the years immediately following discovery and although growth subsequently slows, it is still continuing some 80 years later. Such findings are typical for US 1P data: for example, Lore et al. [76] found that the estimated size of offshore fields in the Gulf of Mexico doubled within six years of discovery and quadrupled within 40 years, while a later study by Attanasii [77] suggested an eight-fold growth in 50 years.

While many US studies estimate growth functions from the date of field discovery [e.g. 77], much of the development work that contributes to reserve growth occurs after production has commenced — which may be several years later [56]. Hence, several authors estimate growth functions from the date of first production [66,67,78]. Both approaches use age as the sole explanatory variable for reserve growth and hence neglect time-varying factors such as oil prices which could change the rate of growth by modifying the incentives for development drilling. The importance of economic factors was demonstrated by Forbes and Zampelli [78] who found a strong positive correlation between gas prices and reserve growth in gas fields in the Gulf of Mexico, together with

---

7 The recovery factor is the ratio of the estimated URR to the estimated OOIP. Estimates of regional and global average recovery factors should be treated with caution [47]. For example, Laherrère [48] uses IHS data to estimate a global average recovery factor of only 27%.

8 Exceptions in this intervening period included the work of Hubbert [70], Arps et al. [71], Marsh [72], Pelto [73] and several Canadian studies [74].

---

Please cite this article in press as: Sorrell S, et al., Shaping the global oil peak: A review of the evidence on field sizes, reserve growth, decline rates and depletion rates, Energy (2011), doi:10.1016/j.energy.2011.10.010
a negative correlation between operating costs (as measured by water depth) and reserve growth.

3.3. Variations in reserve growth

Reserve growth may be expected to vary between different regions and between different ages, sizes and types of field. While there is little systematic evidence available, some useful pointers can be obtained from the existing literature. First, reserve growth varies widely between fields within the same region. For example, an analysis of 934 fields in the Gulf of Mexico found that approximately half grew over the period 1975 to 2002, one fifth shrank and the rest showed no significant change [79]. Attansi and Root [53] found that ‘low quality’ (notably heavy oil) fields grew five times more than conventional fields, while a study of 300 US fields showed that significant reserve growth was largely confined to fields with solution gas drive, heavy oils and low permeability in which techniques such as steam injection,

Fig. 5. Reserve growth in oil fields larger than 0.5 Gb in the UKCS. Note: Horizontal axis represents years after first production. Vertical axis is cumulative discoveries for each field, expressed as a percentage of the initial declared reserves. The heavy black line is the simple arithmetic means of percentages, unweighted by volume. This data series is no longer published by the UK government, in part because of inconsistencies in reporting between different operators. Source: BERR.

Fig. 6. Reserve growth in oil fields smaller than 0.5 Gb in the UKCS. Note: Horizontal axis represents years after first production. Vertical axis is cumulative discoveries for each field, expressed as a percentage of the initial declared reserves. The heavy black line is the simple arithmetic means of percentages, unweighted by volume. This data series is no longer published by the UK government, in part because of inconsistencies in reporting between different operators. Source: BERR.

Please cite this article in press as: Sorrell S, et al., Shaping the global oil peak: A review of the evidence on field sizes, reserve growth, decline rates and depletion rates, Energy (2011), doi:10.1016/j.energy.2011.10.010
hydraulic fracturing and CO₂ injection had been employed [80]. This disparity makes the use of regional or global average growth functions problematic. Nevertheless, while the reserve growth for a particular field may be either positive or negative, the cumulative result for large groups of fields is invariably positive with 1P data. Second, reserve growth also varies significantly from one region to another, even when they are geologically similar. For example, studies show significantly greater reserve growth in Norwegian offshore fields than in either UK or Danish fields (Klett and Gautier, 2005; Sem and Ellerman, 1999; Watkins, 2002). Possible explanations for this include differences in field development practices, reserve definitions, reporting practices, treatment of NGLs and economic and regulatory conditions both between countries and over time. Similarly, Verma and Ulmishek [64] found that lack of investment contributed to West Siberian fields growing much slower than US fields of the same size.

Third, the source and extent of reserve growth varies over the life of a field. It seems likely that early stage reserve growth is more influenced by growth in OOIP while later stage growth is more influenced by changes in recovery factors [81]. Growth is frequently very rapid immediately after field discovery reflecting continuing delineation of the reservoirs, but once production is well-established growth derives more from implementation of EOR, optimisation of well spacing and improved understanding of reservoir characteristics (Verma and Ulmishek, 2003).

Fourth, large fields grow more than small fields (see Figs. 5 and 6). For example, Verma and Ulmishek [64] found that Siberian fields with a URR exceeding 1 Gb doubled in size in 19 years, while smaller fields increased by only 19% over the same period. The production-weighted average for all fields was a 95% increase, since larger fields dominate total reserve additions. Similarly, Grace [79] found that growing fields contained 80% of the discovered hydrocarbons in the Gulf of Mexico and were on average six times larger than fields that shrank. One possible explanation is that smaller fields are more completely explored before the confirmation of reserves, leaving less scope for growth in the estimated OOIP, or that reserve reporting practices differ between large and small fields. But Grace also found that the dominant mechanism of reserve growth was the discovery of new reservoirs which is more likely to occur in large fields. These results suggest that reserve growth may decline in the future (in both absolute and percentage terms) as the average size of new discoveries declines. However, other studies have found no statistically significant correlation between field size and reserve growth [57,78].

Finally, some evidence suggests that onshore fields grow by more than offshore fields [66] and older fields grow by a greater proportion than more recent discoveries. For example, Forbes and Zampelli [78] find a shift to a lower ‘reserve growth regime’ in the Gulf of Mexico after 1987, following the more widespread use of 3-D seismic techniques that allowed more accurate estimation of the size of newly discovered fields. Again, these results suggest that reserve growth may decline in the future as a greater share of production derives from newer fields that are more likely to be located offshore.

3.4. Estimating global reserve growth

Most analysis of reserve growth has taken place in the US, where regulatory rules require the reporting of a particularly conservative interpretation of 1P reserves that is confined to oil that is in contact with a well. As a result, authors such as Laherrère [82] argue that the primary source of observed reserve growth is conservative reporting.9 In contrast, ‘optimists’ such as Mills [6] highlight the historic and potential contribution of improved technology.

This disagreement was brought to a head by the publication of the USGS (US Geological Survey) World Petroleum Assessment in 2000 which provided an authoritative estimate of the global URR of conventional oil. The USGS had previously considered that cumulative 2P discovery data provided a reasonable estimate of the URR of discovered fields, but a growing body of evidence indicated that this assumption was incorrect. For example, the cumulative 2P discoveries for 186 giant fields outside the US were found to have increased by 26% between 1981 and 1996 [83]. The 2000 study therefore included explicit allowance for reserve growth for the first time. The USGS multiplied the cumulative 2P discovery estimates for non-US fields by growth factors that depended upon the age of the field. The latter in turn were derived from a growth function estimated from US data on cumulative 1P discoveries, owing to the lack of adequate data from other regions of the world [84,85]. This process added 654 Gb to the mean estimate of the global URR which was equivalent in size to the estimated yet-to-find resources.10

The USGS acknowledged the limitations of this approach and assumed a triangular probability distribution to account for the associated uncertainty. The application of a 1P growth function to 2P data seems likely to overestimate recoverable resources, but subsequent examination suggests that the USGS assumptions have proved remarkably accurate. For example, Klett et al. [86] found that a total of 171 Gb had been added through reserve growth at

---

9 Reserve growth in the US may also be influenced by the production restrictions imposed in the 1950s and 60s (‘pro-ratoning’) and by the underestimation of field size in the early days of exploration owing to inferior geophysical techniques.

10 This corresponds to a 44% growth in the estimated URR of crude oil fields discovered before 1995 and a corresponding 56% growth in NGL resources.
non-US fields between 1995 and 2003, or more than twice the reserve additions through new discoveries. This suggests that 28% of the mean USGS estimate for non-US reserve growth had been added to in the first 27% of the assessment time frame (1995–2025). Similarly, Stark and Chew [87] found a global total of 465 Gb of reserve growth between 1995 and 2003, of which 175 Gb was attributed to ‘classic’ reserve growth and the remainder to ‘new and revised data’. This distinction suggests that much of the apparent reserve growth could derive from factors such as the inclusion of previously omitted fields in the industry databases and from revised estimates of fields where the data was poor. The biggest growth in absolute terms derived from Middle East fields where the reserves data is particularly uncertain.

To check whether the rate of reserve growth observed by Klett et al. is being maintained, we compared the 2000 and 2007 iterations of the IHS Energy PEPS database. This provides country-level data on oil production and reserves and backdates any reserve revisions to the year in which the relevant fields were discovered — thereby providing a more accurate indication of what was ‘actually' found at that time. With backdating, estimates of the volume of discoveries made in a given year are constantly being revised.

The results (Fig. 8) suggest that cumulative 2P discoveries for pre-2000 fields grew by 11% between 2000 and 2007. The global figure includes US and Canadian data which is not comparable with the rest of the database since it is based upon 1P reserves. If the US data is removed, pre-2000 fields are estimated to have grown by 13.9% between 2000 and 2007, suggesting that the rate of non-US reserve growth has increased in recent years. The percentage reserve growth varies widely from one country to another with the largest contribution in absolute terms deriving from Saudi Arabia and Iran (Fig. 9). Interestingly, most of the growth derives from fields discovered before 1986, with very little growth in the more recently discovered fields.

In summary, though the actual rates may be contested, it is clear that significant reserve growth is observed in cumulative discovery estimates based upon both 1P and 2P reserves. With 2P data, the global average reserve growth observed since 1995 seems broadly in line with the crude and controversial assumptions made by the USGS [83]. However, the global average is strongly influenced by reserve growth in countries with the largest reserves where there is less confidence in the accuracy of the data. Also, it is far from clear that the observed trend in reserve growth will be maintained in the future. Reserve growth appears to be greater for larger, older and onshore fields, so as global production shifts towards newer, smaller and offshore fields the rate of reserve growth may decrease in both percentage and absolute terms. At the same time, higher oil prices may stimulate the more widespread use of EOR techniques that have the potential to substantially increase global reserves. The suitability of such techniques for different types of field and the rate at which they may be applied remain key areas of uncertainty.

4. Decline rates

An critical determinant of future investment needs and supply trends is the rate of decline of production from currently producing fields. Supply forecasts are more sensitive to assumptions about the rate of decline than to assumptions about future oil demand, but the former have generated controversy owing to lack of data [88]. Fortunately, three recent studies have put a great deal of data into the public domain. This section examines the nature of production decline, summarises evidence on the rate of decline from different categories of field, compares estimates of global average decline rates and highlights the implications of increasing decline rates for future supply.

4.1. Analysis of production decline

The production cycle of individual fields can vary widely depending upon their geology and location and the manner in which they are developed (Fig. 10). As a field is brought on-line, its rate of production typically rises rapidly to a peak which may extend into a multi-year plateau as a consequence of the limited capacity of pipelines and other surface facilities and/or the steady development of the field through additional drilling. The length of plateau tends to be greater for large fields and the production cycle can be complicated by interruptions and the introduction of new technology. But at some point, the rate of production will begin to decline as a result of falling pressure and/or the breakthrough of water (in mature fields, the ‘water-cut’ may represent 90% or more of the volume of produced liquids.) Typically, more than half of the recoverable resources of a field will be produced during the decline phase.

The term ‘decline’ is loosely applied at various levels of aggregation, including single wells, reservoirs, fields, basins and countries. When applied to a region, it is important to distinguish between the overall decline rate which refers to all currently producing fields, including those that have yet to pass their peak, and the post-peak decline rate which refers to the subset of fields that are in decline. Since the production cycle of individual fields is rarely smooth, the point at which decline begins can be ambiguous. Some analysts also estimate natural decline rates which exclude the effects of capital investment.

Production from individual wells, reservoirs and fields is usually assumed to decline exponentially at a constant rate, although there is no physical law requiring this and the rate of decline often falls during the later stages of the production cycle. Empirical equations to model production decline were first developed over a century ago and have since seen wide application (Box 2) [89–91]. The exponential model is the most widely used, but it can underestimate production during the later stages of a field’s life.

As an illustration, Fig. 11 shows production profiles for the UK’s three largest offshore fields (Forties, Brent and Ninian) and its largest onshore field (Wytch Farm). Forties can be approximated with a 9%/year exponential decline, Ninian by 11%/year, and Brent a poor fit of around 13%. Wytch Farm’s decline rate is similar to the offshore fields, but more oil was produced pre-peak. Fig. 12

Please cite this article in press as: Sorrell S, et al., Shaping the global oil peak: A review of the evidence on field sizes, reserve growth, decline rates and depletion rates, Energy (2011), doi:10.1016/j.energy.2011.10.010
shows the annual production for a group of 77 UKCS fields which peaked in or before 1996. This gives an estimate of \( \sim 12.5\%/year \) for the aggregate post-peak decline rate. At least 60% of cumulative production occurred during the decline phase and these fields are still producing. Fig. 13 suggests that younger UKCS fields have steeper decline rates, but since these fields are also smaller on average, the observed trend relates to both the size and age of the field.

4.2. Regional and global average decline rates

Three studies have estimated decline rates from a globally representative sample of fields (Table 2). While each study uses a different sample, they all include the giant fields which account for around half of global production. Unfortunately, the studies use competing definitions and different approaches to production weighting.

These studies estimate the production-weighted decline rate of their sample of post-peak fields to be 5.1%/year (IEA), 5.5%/year (Höök et al.) and 5.8%/year (CERA) (Table 3). The production-weighted decline is less than the average decline because fields with higher production tend to be larger and decline more slowly. The studies also agree that:

- Decline rates are lower for OPEC fields and particularly for Middle East fields (Table 3). This is partly reflects differences in average size, but also quota restrictions and disruptions from political conflict.
- Decline rates are higher for offshore fields (Table 3). These tend to be produced at higher rates in order to recover their higher fixed costs, leading to higher peaks, shorter plateaus and steeper declines.
- Decline rates are lower for larger fields and are particularly low for the super-giant fields in the Middle East (Table 4). Large fields reach their peak later than small fields, but also produce a greater proportion of their URR during the decline phase.

Importantly, both the IEA and Höök et al. find decline rates to be significantly higher for newer fields (Fig. 14). The IEA argues that newer fields build-up more quickly to a higher plateau that is maintained over a shorter period of time, but Höök et al. [92] show that the length of plateau for giant fields has increased together with the proportion of the remaining recoverable resources produced prior to peak. They argue that new technology allows the plateau to be maintained for extended periods of time, but at the cost of more rapid decline following the peak [see also Ref. [93]].

The above figures are likely to underestimate the global average decline rate for all post-peak fields since the mean size of each sample of fields is greater than that of the global population. Under the optimistic assumption that that decline rate for smaller fields is the same as that for the sample of large fields (10.4%), the IEA estimate a production-weighted global average decline rate of \( 6.7\%/year \) for all post-peak fields. With capital investment reduced as a result of the 2008 economic recession, the average decline rate may have subsequently increased.

This decline in production needs to be replaced by investment in EOR at producing fields, the development of 'fallow' fields or the discovery and development of new fields. But to estimate the additional capacity required each year it is necessary to know either the proportion of production from post-peak fields, or the production-weighted aggregate decline rate of all fields, including those in build-up. Both figures are absent from the IEA study and...
Box 2. Empirical equations to model production decline

Production decline from oil wells was first modelled by Arnold and Anderson [105] and subsequently by Cutler [106] and Larkey [107] among others. These early studies were consistent with primary recovery being driven by the expansion of natural gas. Contemporary decline curve analysis has its roots in Arps [108], who introduced empirical curves defined by three variables: the initial rate of production \( Q'(t_0) \), the curvature of decline \( \beta \) and the rate of decline \( l \).

The general hyperbolic equation for the rate of production is:

\[
Q'(t) = \frac{Q'(t_0)}{\left(1 + \beta \left(t - t_0\right)\right)^{1/\beta}}
\]

If \( \beta = 0 \), this reduces to exponential decline:

\[
Q'(t) = Q'(t_0) e^{-l\left(t - t_0\right)}
\]

If \( \beta = 1 \), this reduces to harmonic decline:

\[
Q'(t) = \frac{Q'(t_0)}{\left(1 + \lambda \left(t - t_0\right)\right)}
\]

Decline models have since been developed in a variety of ways, including linearised curves [109–111], and the econometric analysis of residuals [112]. Kemp and Kasim [113] found that a logistic curve provided a better fit for UKCS fields. Decline curves are commonly used to estimate the URR of a field (Section 4.1).

they appear to calculate the capacity requirements incorrectly.\(^{11}\) We estimate the latter figure to be \(-4.1\%/\text{year}\) which compares to CERA’s estimate of \(4.5\%/\text{year}\). This implies that at least \(3\ \text{mb/d}\) of capacity must be added by new investment each year, simply to maintain production at current levels — equivalent to a new Saudi Arabia coming on stream every three years. If demand grows and/or decline rates increase, larger volumes of capacity will be required.

A critical question for supply forecasting is how global average decline rates may be expected to develop in the period to 2030. Most existing fields will enter decline over this period, with

\[\text{Fig. 11. Production from four UK oil fields fitted by three exponential decline models. Source: UK Department of Energy and Climate Change.}\]

\[\text{Fig. 12. Production from UK offshore fields which peaked before 1997, stacked by peak year. Source: UK Department of Energy and Climate Change.}\]


\(^{11}\) The IEA appear to multiply the production-weighted decline rate of post-peak fields (6.7%) by global crude production (70 mb/d) to estimate an annual loss of output of 4.7 mb/d. But the correct procedure is to use the production-weighted aggregate decline rate of all fields, including those in build-up. The IEA provide this figure for OPEC (3.3%) and non-OPEC fields (4.7%) so we simply weight by 2007 production to obtain a global average. The result appears consistent with IEA’s graphs (Fig. 15).
a growing proportion of production from younger, smaller, and offshore fields that tend to have higher fixed costs and hence higher rates of post-peak decline. The IEA [1] anticipates the production-weighted global average decline rate of post-peak fields to increase to 8.5%/year by 2030, leading to an estimated loss of 61% of current capacity (43 mb/d) (Fig. 15). However, Höök et al. [92] consider this estimate to be optimistic, given the trend towards increasing decline rates in the giant fields and the observed tendency for the production-weighted decline rate to converge on the (higher) average rate [94].

In summary, the global average decline rate of post-peak fields is at least 6.5%/year and the corresponding decline rate of all currently producing fields is at least 4%/year. Both are on an upward trend as more giant fields enter decline, as production shifts towards smaller, younger and offshore fields and as changing production methods lead to more rapid post-peak decline. Significant investment is needed simply to offset the underlying natural decline rates and if this is not forthcoming decline rates will increase. While future trends in decline rates are difficult to forecast, a case could be made that the IEA’s assumptions are optimistic. If so, more than two thirds of current crude oil production capacity will need to be replaced by 2030, simply to keep production constant. Given the long-term decline in new discoveries, this will present a major challenge even if ‘above-ground’ conditions prove favourable.

5. Depletion rates

Decline rates are a measure of the change in the rate of production of a field from one year to the next. They should not be confused with depletion rates which are a measure of the rate at which the recoverable resources of a field or region are being produced. The depletion rate of individual field is defined as the ratio of annual production to some estimate of recoverable resources, where the latter could be the 1P reserves, the remaining recoverable resources (i.e. allowing for future reserve growth) or the URR. When defined in relation to 1P reserves, the depletion rate is simply the inverse of the more familiar reserve to production (R/P) ratio. While decline rates can be measured precisely, depletion rates are based upon uncertain resource estimates which vary between sources and over time – with higher resource estimates leading to lower estimates of depletion rates.

Höök [95] defines the depletion rate as the ratio of annual production to remaining recoverable resources, where the latter is calculated by subtracting cumulative production from an estimate of the URR. He then shows the close links between the depletion rate and the decline rate of a field. The depletion rate generally increases during the build-up and plateau phase as reserves are produced. Once decline begins, the depletion rate either remains constant or falls – provided the URR estimate remains unchanged. If decline is exponential the depletion rate equals the decline rate, while if decline is hyperbolic the maximum depletion rate is reached just prior to the onset of decline. Höök et al. [27] show that the maximum depletion rate of giant oil fields typically falls within a relatively narrow band, with a production-weighted mean of 7.2%/year (Table 5). As with decline rates, the maximum observed depletion rate is higher for offshore fields and lower for OPEC fields.

Höök et al. [27] also estimate the depletion at peak, or the proportion of URR produced at the onset of field decline. The production-weighted mean of their sample of giant fields is 37%, with the average being higher for offshore fields and lower for OPEC fields. A similar analysis is conducted by the IEA [1] who find values ranging from 15% for large fields to 25% for small offshore fields. These results demonstrate that production from most fields begins to decline when less than half (and often less than one third) of their recoverable resources have been produced. The IEA estimates that giant fields are on average 48% depleted, with the regional average varying from 37% in the Middle East to 78% in North America.

Table 2
Comparison of global decline rate studies.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>IEA</th>
<th>Höök et al.</th>
<th>CERA</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of fields in sample</td>
<td>651 (54 super-giant, 263 giant, 334 large)</td>
<td>331 (all giant)</td>
<td>811 (400 large and above)</td>
</tr>
<tr>
<td>No. post-peak fields</td>
<td>580&lt;sup&gt;a&lt;/sup&gt;</td>
<td>291&lt;sup&gt;b&lt;/sup&gt;</td>
<td>–</td>
</tr>
<tr>
<td>% of total production of crude oil in sample</td>
<td>~58%</td>
<td>~50%</td>
<td>~66%</td>
</tr>
<tr>
<td>Cumulative discoveries of crude oil in sample</td>
<td>1241 Gb</td>
<td>1130 Gb</td>
<td>1155 Gb</td>
</tr>
<tr>
<td>Definition of plateau</td>
<td>Production &gt;85% of peak</td>
<td>Production &gt;96% of peak</td>
<td>Production &gt;80% of peak</td>
</tr>
<tr>
<td>Definition of onset of decline</td>
<td>After year of peak production</td>
<td>After last year of plateau</td>
<td>After last year of plateau</td>
</tr>
<tr>
<td>Production weighting</td>
<td>Cumulative production&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Annual production</td>
<td>Annual production</td>
</tr>
</tbody>
</table>

Notes:
- <sup>a</sup> 101 fields in plateau (production >85% of peak), 117 fields in ‘phase 1 decline’ (production >50% of peak), 362 fields in ‘phase 3’ decline (production <50% of peak).
- <sup>b</sup> 387 onshore, 264 offshore, 185 OPEC and 466 non-OPEC.
- <sup>c</sup> 261 onshore, 214 offshore, 143 OPEC and 188 non-OPEC.

Table 3
Estimates of production-weighted aggregate decline rates for samples of large post-peak fields (%/year).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>IEA</th>
<th>Höök et al.</th>
<th>CERA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Onshore</td>
<td>4.3</td>
<td>3.9</td>
<td>–</td>
</tr>
<tr>
<td>Offshore</td>
<td>7.3</td>
<td>9.7</td>
<td>–</td>
</tr>
<tr>
<td>Non-OPEC</td>
<td>7.1</td>
<td>7.1</td>
<td>–</td>
</tr>
<tr>
<td>OPEC</td>
<td>3.1</td>
<td>3.4</td>
<td>–</td>
</tr>
<tr>
<td>Total</td>
<td>5.1</td>
<td>5.5</td>
<td>5.8</td>
</tr>
</tbody>
</table>

Note: Studies use different data sets, definitions and methods of production weighting. Details missing for CERA since we do not have access to the full study. Source: IEA [1], CERA [114] and Höök et al. [27,92,94,95].

Table 4
IEA estimates of aggregate production-weighted decline rates for different sizes of post-peak fields (%/year).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Total</th>
<th>Super-giant</th>
<th>Giant</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offshore</td>
<td>4.3</td>
<td>3.4</td>
<td>5.6</td>
<td>8.8</td>
</tr>
<tr>
<td>Offshore</td>
<td>7.3</td>
<td>3.4</td>
<td>8.6</td>
<td>11.6</td>
</tr>
<tr>
<td>Non-OPEC</td>
<td>7.1</td>
<td>5.7</td>
<td>6.9</td>
<td>10.3</td>
</tr>
<tr>
<td>OPEC</td>
<td>3.1</td>
<td>2.3</td>
<td>5.4</td>
<td>9.1</td>
</tr>
<tr>
<td>All fields</td>
<td>5.1</td>
<td>3.4</td>
<td>6.5</td>
<td>10.4</td>
</tr>
</tbody>
</table>

Note: The production-weighted decline rate is 1.4% in decline phase 1, 3.6% in decline phase 2 and 6.7% in decline phase 3. The production-weighted average for phase 1 is strongly influenced by Ghawar. The production-weighted sample average for post-plateau fields is 5.8%. Source: IEA [1].
Depletion and depletion rates can also be estimated at the regional level, although the uncertainty on the recoverable resource estimates will necessarily be greater since they also include estimates of the yet-to-find resources. Of particular interest are the values at peak for the countries that have passed their peak of production. Using USGS estimates of regional URR for 37 post-peak countries, we estimate a simple mean for depletion at peak of 22%, a production-weighted mean of 24% and a maximum of 52% (Fig. 16). In other words, most countries appear to have reached their peak well before half of their recoverable resources have been produced.

In a similar manner, we estimate the mean depletion rate at peak for these countries to be 2.1%/year, the production-weighted mean to be 1.9%/year and the maximum to be 5.2%/year (for Bulgaria which is an outlier). In other words, the maximum depletion rate for a region has typically been much less than 5%/year. Also, the average depletion rate over the full production cycle is typically much lower than the maximum rate. At present, the global average depletion rate is approximately 1.2%.

This analysis suggests that there are physical, technical and economic constraints on both the rate of depletion of a field or region and the proportion of the URR that can be produced prior to the peak. Hence, both measures can provide a useful ‘reality check’ on regional and global supply forecasts — although the usefulness of these ‘rules-of-thumb’ depends very much upon the accuracy of the estimates of resource size. Specifically, any forecast that implies depletion rates that are significantly higher than those previously experienced in oil-producing regions will require careful justification. The same applies to forecasts that delay regional peaks of production until significantly more than half of the URR of that region has been produced.

These points are especially relevant to the influential IEA World Energy Outlook [1]. The reference scenario of the 2008 edition projects 114 Gb of new discoveries by 2030 and anticipates these fields producing 19 mb/d by 2030 with cumulative production of 46 Gb [3]. As Aleklett et al. [3] demonstrate, this implies an average depletion rate of 10% (and rising) from newly discovered resources by 2030, which far exceeds the historical experience of any oil-producing region anywhere in the world. A comparable analysis for fallow fields suggests even higher average depletion rates of 12–13% (and rising) by 2030. These figures appear implausible, and raise the suspicion that the IEA worked backwards from what these fields were required to produce in order to meet projected demand, rather than forwards from what production rates were considered reasonable. But despite the importance of this critique, the IEA has yet to provide any justification for their assumptions.

Depletion rates can also provide a useful bridge between estimates of the rate of reserve growth and/or new discoveries and the rate of production. While it is common to estimate reserve additions in Gb/year, to translate this into a feasible rate of production it is necessary to multiply by an assumed depletion rate. If the product of the two is less than the capacity anticipated to be lost through production decline, then aggregate production in a region may be expected to fall. For example, a global average decline rate of 4.1% implies an annual loss of 2.9 mb/d or ~1.0 Gb/year of production capacity. This capacity needs to be replaced by a combination of developing fallow fields, reserve growth at existing fields and new discoveries simply to maintain production at current levels. Using a peak depletion rate of ~5.0%/year, this leads to a requirement for ~20 Gb/year of reserve additions from these sources if global production is to be maintained. If instead the depletion rate of these resources is only 1.2%/year (the current global average for all production), reserve additions of ~80 Gb/year are required. These figures compare to global reserve additions of approximately 45 Gb/year between 2002 and 2007, two thirds of which derived from reserve growth. As demand grows and decline rates increase in the medium to long-term, the rate at which reserves are added from these sources, or the rate at which they are depleted needs to increase. However, the former runs counter to both the long-run trend of declining discoveries and the anticipated decline in the contribution from reserve growth (Section 3), while the latter is subject to physical, engineering and economic limits.

6. Summary

This paper has reviewed the evidence on field-size distributions, reserve growth, decline rates and depletion rates for crude oil. It is

<p>| Table 5 Estimated depletion at peak and annual depletion rate at peak for giant oil fields. |</p>
<table>
<thead>
<tr>
<th>Depletion at peak</th>
<th>Depletion rate at peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Onshore</td>
<td>34.1%</td>
</tr>
<tr>
<td>Offshore</td>
<td>44.0%</td>
</tr>
<tr>
<td>Non-OPEC</td>
<td>37.4%</td>
</tr>
<tr>
<td>OPEC</td>
<td>31.5%</td>
</tr>
<tr>
<td>All fields</td>
<td>36.8%</td>
</tr>
</tbody>
</table>

Notes:
- Depletion rate = Ratio of annual production to estimated remaining recoverable resources.
- Depletion = Ratio of cumulative production to estimated ultimately recoverable resources. All figures production-weighted. Source: Höök et al. [27].

---

12 We estimate the depletion rate at peak for the UK to be 4.4% assuming a URR of 43 Gb. In contrast, Aleklett et al. [3] estimate a depletion rate at peak of 6.9% assuming a lower URR of 35 Gb. This discrepancy highlights the sensitivity of these estimates to the assumed URR.
Fig. 16. Peak production, peak year and depletion at peak for 37 post-peak countries. Note: Shows peak year and estimated percentage of USGS (2000) estimate of URR that was produced by the date of peak. Since the USGS only estimate reserve growth at the global level, this was allocated between countries in proportion to their estimated URR excluding reserve growth. Timing of the peak of production for many of these countries is influenced by factors other than physical depletion and in some cases the peak may subsequently be exceeded. Source: BP (2008); USGS (2000).
shown that, despite serious data limitations, the level of knowledge of each of these issues has improved considerably over the past decade. While the evidence on reserve growth appears relatively encouraging for future global oil supply, that on decline and depletion rates does not. The main findings are as follows:

Around 100 oil fields account for up to half of the global production of crude oil, while up to 500 fields account for two thirds of cumulative discoveries. Most of these fields are relatively old, many are well past their peak of production and most of the rest will begin to decline within the next decade or so. The remaining reserves at these fields, their future production profile and the potential for reserve growth is therefore of critical importance for future global supply.

While the observed lognormal size distribution of discovered fields is partly the result of sampling bias, there is insufficient evidence to conclude whether a ‘linear’ or ‘parabolic fractal’ better describes the population size distribution. Moreover, while technical improvements and higher prices should make more small fields viable, many of the smallest will remain uneconomic to develop and the exploitation of the rest will be subject to rapidly diminishing returns. As a result, the resources contained in small, undiscovered fields are of much less significance to future supply than the potential for increased recovery from the giant fields.

Reserve growth is real, significant and not primarily the result of conservative reserve reporting. Increased recovery from known fields should therefore make a major contribution to future global oil supply. The global average reserve growth observed since 1995 seems broadly in line with the optimistic assumptions made by the USGS [83], but this figure is strongly influenced by reserve growth in countries with both the largest reserves and the poorest quality data. Also, reserve growth appears to be greater for larger, older and onshore fields, so as global production shifts towards newer, smaller and offshore fields the rate of reserve growth seems likely to decrease in both percentage and absolute terms.

A critical determinant of investment needs and future supply is the rate of decline of production from existing fields. The production-weighted global average decline rate of post-peak fields is at least 6.5%/year and the corresponding decline rate of all currently producing fields is at least 4%/year. Both are on an upward trend as more giant fields enter decline, as production shifts towards smaller, younger and offshore fields and as changing production methods lead to more rapid post-peak decline. More than two thirds of current crude oil production capacity may need to be replaced by 2030, simply to keep production constant. At best, this is likely to prove extremely challenging.

There are physical, technical and economic constraints upon the rate of depletion of a field or region. At the regional level, the maximum observed depletion rate is ~5%/year, and the average depletion rate over the production cycle is typically much less. Also, most giant fields and most countries appear to have reached their peak well before half of their recoverable resources have been produced. Hence, supply forecasts that assume or imply either higher depletion rates or the production of more than half the recoverable resources prior to the peak require careful justification.

At present, the reference scenario in the IEA World Energy Outlook appears inconsistent with historical depletion rates and seems to rely upon optimistic assumptions about future decline rates. Since slower depletion or more rapid decline implies an earlier global peak, these assumptions deserve serious investigation.

Acknowledgements

This work was funded by the UK Energy Research Centre (UKERC) as part of a wide-ranging study of global oil depletion [7]. The authors would like to thank IHS Energy for allowing the publication of data from their PEPS database, Robert Kaufmann and Jean Laherrère for their helpful comments and Rob Gross, Phil Heptonstall and Jim Skea for their support. The usual disclaimers apply.

References


Please cite this paper in press as: Sorrell S, et al., Shaping the global oil peak: A review of the evidence on field sizes, reserve growth, decline rates and depletion rates, Energy (2011), doi:10.1016/j.energy.2011.10.010
Klett TR, Wang PPC. Probabilistic formulation of a method for the evaluation of petroleum resources. Journal of the International Association of Mathemat-
Power M. Lognormality in the observed size distribution of oil and gas pools as a consequence of sampling bias. Mathematical Geology 1992;24.
Lee PJ, Wang PPC. Probabilistic formulation of a method for the evaluation of petroleum resources. Journal of the International Association of Mathemat-
Power M. Lognormality in the observed size distribution of oil and gas pools as a consequence of sampling bias. Mathematical Geology 1992;24.
Lee PJ, Wang PPC. Probabilistic formulation of a method for the evaluation of petroleum resources. Journal of the International Association of Mathemat-
Power M. Lognormality in the observed size distribution of oil and gas pools as a consequence of sampling bias. Mathematical Geology 1992;24.


[106] Cutler WC. Estimation of underground oil reserves by oil-well production curves. Department of the Interior; 1924.


[111] Luther LC. Linearization and regression analysis technique predicts hyperbolic decline in reserves. Oil and Gas Journal; 1985:78–9.

