Measuring Evidence Uptake

December 2019

Executive summary

Nesta's innovation mapping team seeks to support innovation policymakers by supplying them with evidence to inform their decisions. In addition to developing new data collection, processing and analysis techniques, Nesta delivers evidence through a novel channel as well: self-service platforms that enable users to explore the data directly through interactive visualisations. The innovation mapping team is presently looking to develop a system to measure the uptake of the evidence it produces, to understand how users are engaging with it and ultimately to be able to offer even better support to innovation policymakers. For this purpose, they contracted The Decision Lab to map out established and emerging approaches to measuring the use of evidence in policy, as an input to guide Nesta in developing and pilot-testing its own system.

The resulting report – which has here been edited for length and to make it more applicable to a wider audience – takes note of the current context of supplying evidence for policy, a context in which the diversity of suppliers and data types has grown enormously in recent years and in which the need to demonstrate 'impact' is a key driver of behaviour. Engaging critically with this context and noting four tensions that measurement systems face as a result, this report puts forward a number of concrete suggestions:

1. Acknowledging that the policy ecosystem involves several interwoven threads, evidence providers should identify the policymakers that they seek to support and identify where records of their discussions and decisions can be accessed. Furthermore, acknowledging that public discourse is an important component in policy deliberation, they should identify which 'publics' are relevant to their work and identify where traces of their discourse can be collected as well; both news and social media are highlighted here. With these data sources identified, an automated system could be developed to harvest the output through these various channels, for subsequent analysis.
2. Acknowledging that evidence is packaged in many different ways as it travels through these channels, evidence providers should put together an inventory of the different ways in which their evidence might be represented; examples include peer-reviewed papers, data sets, blogs/blog posts, podcasts/podcast episodes, interviews, social media posts and so forth. With this taxonomy developed and populated, they could then use this inventory as a list of targets to seek in the data sources noted above.
3. Acknowledging that evidence can play various roles within the policy discussion and that a dichotomy between impact/no impact is harshly reductive, evidence providers should develop a taxonomy of types of consideration that they consider relevant to their operations; instrumental use and conceptual use are noted here as the basis for many discussions. With this taxonomy in hand, they could develop a system to tag mentions of the evidence – in various packages, on various channels – by the type of consideration that it exhibited in an individual case.
4. Recognising the increasingly important role played by online platforms in the provision of data and analyses, evidence providers should develop a set of customer journey maps, user personas and a set of touchpoints with their platform(s) and indicators should be chosen to measure these. Using these indicators makes it possible to measure user behaviour against the personas and journeys that have been developed – which are hypotheses and therefore call for testing and iterative refinement.
5. Acknowledging the value of connecting usage of the platform to the consideration of evidence within the wider policy setting, evidence providers should develop a way to identify which types of platform behaviours promote (or at least correlate with) wider diffusion of that evidence. For this purpose, two existing tools are identified – SEER and the theory of planned behaviour – along with the possibility of simplifying and integrating them into a short user survey.

In addition to these practical outputs, the report also engages substantively with the theoretical and even philosophical considerations about evidence-based policymaking. This engagement with theory underpins the measurement recommendations put forward. The report therefore attempts also to make a substantive contribution to theoretical perspectives on the evidence–policy interface.

1. Introduction

1.1: The changing landscape of evidence for policy

Traditionally, national statistical agencies have been major providers of evidence into government decision-making processes. The data that they produce are often collated from administrative sources within the government (e.g., statistics aggregated across tax filings) or collected specifically for the purpose of statistical measurement (e.g., a periodic census of the population).

In both cases, the positioning of the agency – being located within the government – has been crucial to its function: providing the agency with access to confidential documents within government, mandates to legitimise solicitation of members of the public, the computing resources required to work with the raw data and so forth. Other government departments as well as academic institutions have also played important roles in the provision of evidence to decision-makers, similarly benefiting from the access, legitimacy and resources of the public sector.

However, with the major developments in information communication technologies over recent decades, the situation has shifted considerably. Much of our day-to-day activity has transitioned from the analogue to the digital world, leaving a larger data footprint. Those interactions are also carried out with a wide variety of actors, located across the sectors of society, including various layers of intermediaries such as online platforms. Our data footprints are therefore both larger and more widely spread out.

The dramatic cost reductions in digital storage space and computing capacity have enabled these digital footprints to be collected, collated and analysed by a growing range of players. In short, novel systems for data collection, processing, analysis, visualisation and delivery have emerged and these are in the hands of a wide range of actors within society. National statistical agencies (along with further government departments and academic institutions) remain important sources of evidence, especially in a government decision-making context. However, the field is much more diverse and much more crowded now than it was even a few years ago.

Actors within the evidence advisory landscape wish to better understand how their evidence is used to support policy. For this purpose, it will be necessary to experiment with a system to measure the uptake of the evidence. Such systems add another layer to the evolution of evidence for policy; new evidence-development systems are emerging, along with systems to gather evidence about the use of that very evidence. Thus, there are new sources to inform us about the functioning of many ecosystems for which we build policy and new sources to inform us about the functioning of the policy ecosystem itself.

1.2: The impact agenda

Much of the discourse around measuring evidence uptake is driven by the 'impact agenda.' In the aftermath of the Great Recession, there has been increased pressure on government spending, which has led to an increased urgency for governments (as well as other bodies receiving public support, such as academic institutions) to demonstrate their impact as the 'return' that they deliver on public 'investment.' The pressures to measure and demonstrate impact have come to be known as the 'impact agenda.' Intense discussion has followed in many corners about the effect that these pressures are having on the public sector and beyond, with some referring to this situation as the 'tyranny of the audit' (Bossuyt, Shaxson and Datta, 2013).

Governments seek to have impact on society through policy, whereas researchers seek to have impact on policy through the uptake of their evidence. In this way, the measurement of policy impact has been markedly shaped by the wider discourse around impact; there are wide coalitions within the research sphere around responsible ways to develop and use metrics (American Society for Cell Biology, 2013; Hicks, Wouters, Waltman, de Rijcke and Rafols, 2015; Wilsdon et al., 2015). Understanding this wider 'impact' context provides important considerations for reflecting on the findings of the present study and selecting a path forward in measuring how their evidence supports policy. A brief overview of the impact agenda is therefore valuable here.

One of the key tensions that arises in this context is between evidence for accountability purposes (where central authorities verify that public funds have delivered relevant impacts, to decide about programme budgeting) as opposed to evidence for learning purposes (where local programme managers examine how the programme functions to deliver its value, to decide how to improve its operation). Given the accentuated pressure to demonstrate return on investment, any hint that there is room for improvement comes with an increased risk that this will be interpreted as inefficiency and thus waste, ultimately leading to the programming budget being cut. If we constantly diminish possibilities for improvement, we make it challenging to learn; if the program is already optimal (as we constantly reaffirm), then things ought to remain as they are. The focus on accountability thus effectively marginalises the possibility for learning.

1.3: Piloting a system for learning

Evidence of policy uptake is in many cases collected under the aegis of the impact agenda and as we have just explored, that agenda is oriented primarily towards accountability rather than learning. If the priority in measuring uptake is to facilitate learning rather than accountability, in order to provide even better support to policymakers rather than to demonstrate that they are already providing (optimal) support, much of the existing machinery to measure uptake must be repurposed if it is to be used here. Accordingly, the task with which The Decision Lab was mandated had to include both a summarising function and a critical function, identifying major trends while seeking to understand how steering forces have driven those trends in specific directions.

The present report aims to identify and critically assess approaches to measuring the uptake of evidence in policy, with the aim of ultimately being applied in practice. Thus, this report aims to provide options that are well-grounded in theory, that are actionable and that are forward-looking (acknowledging the rapid evolution both of the evidence-for-policy ecosystem and the approaches to measuring that ecosystem).

1.4: Structure of the report

The report is structured in such a way as to facilitate the wider uptake of the findings. The introduction sets the conceptual context in which the main report is situated.

Two very simple measurement approaches are described below: bibliometric measurement and first-hand reports (gathered through surveys or interviews). These are the approaches to evidence-uptake measurement that one most frequently encounters. Because these ubiquitous approaches occupy such a vast territory within this space, other approaches from the literature are presented by way of contrast to them.

The present report therefore adopts something of a similar stylistic approach. The two 'baseline' approaches are described briefly below, as the backdrop for the bulk of the discussion. The main body of the report is divided into four subsections, each examining a different shortcoming of the baseline system and its conceptual underpinnings:

1. A linear conception of policymaking is descriptively inaccurate.
2. A linear conception of policymaking is normatively undesirable.
3. The system does not elucidate how users engage with the evidence.
4. The system does not elucidate how engagement promotes consideration of the evidence.

Within each subsection, the text articulates the conceptual or technical deficiency of the simple system. Alternatives that offer a solution to this deficiency are then presented, along with an explanation of how they overcome the deficiency in question. Where more than one alternative was discovered, they are compared to each other. Finally, the subsection ends with a discussion of what would be entailed in practical as well as conceptual terms in adopting such a system.

1.5: The baseline

There are two approaches that dominate the landscape of measuring evidence uptake in policy. The first is to search within policy documents for explicit references to scholarly research outputs. This approach can be operationalised using either a manual approach (with humans reading policy documents looking for references to a list of papers) or an automated approach (with digital systems harvesting policy documents, extracting and parsing references and matching references to a bibliographic database). In either case, let us refer to this as the 'bibliometric' approach.

The second approach is to ask people involved in the policy process (such analysts, decision-makers, or stakeholders) about how the evidence influenced the decision. Again, this approach can be operationalised in a number of ways, such as collecting information through surveys or interviews, for instance. Questions can be about how a certain piece of evidence impacted a decision or about the impact of evidence generally on a decision; they can also be about a specific decision or about the 'normal operation' of decision-making in a given context. What holds these approaches together is that we are asking participants first to reconstruct their mental process (or even the mental processes of others) leading up to a decision and then to assess the role that evidence played in that process. This task that we get participants to perform is often requested long after the decision itself is made rather than while decision-making is in progress. Let us refer to these as the 'rationalisation' approaches.

2. Policymaking is complex

2.1: A challenge for the baseline

The most compelling storylines about impact are those that take place in very linear, deterministic systems. In demonstrating 'impact,' one lays claim to the outcomes (in this case, policy choices), purporting that one's actions have caused these effects to come about, determining a decision either in whole or in part. Attribution therefore becomes a crucial aspect of measuring policy impact, to legitimise one's claim. However, attribution is widely recognised as a notable methodological challenge in measuring impact. In response to calls for greater accountability, it would be a powerful response to be able to demonstrate clearly that 'such and such' evidence led unambiguously to 'such and such' decisions. The appeal of linear narratives creates a pressure to view ecosystems as linear and to create measurement systems built on this premise.

For instance, in their marketing documentation regarding data on policy citations to research, leading providers of alternative research metrics (so-called 'altmetrics') are clearly articulating their offering through the lens of the impact agenda. Announcing the addition of these data to the Dimensions platform (by Digital Science), Christian Herzog is quoted as saying: *"Information on where and how research has been referenced in policy is a critical part of the puzzle for researchers and organizations looking to understand ... real-world impact."*[1] Similarly touting their coverage of policy documents, Plum Analytics (by Elsevier) stated on their blog, *"Finding references to research in policy documents … [enables] researchers and those who support them to demonstrate public engagement with their research, its impact on government policy."*[2] This conceptualisation of citations is imported from the field of bibliometrics, where citations have (despite expressed concerns about their diverse meanings in practice) continued to be treated as an indication of positive influence (Garfield, 1979).

However, public programmes rarely exist in an ecosystem that moves in linear, clearly defined pathways. Carden (2009) describes this iterative complexity nicely, highlighting that *"policy decisions over time generally display a complicated pattern of advances and reversals tied together in feedback loops of decision, implementation, second thoughts and course corrections."*

Additionally, the structures that guide policy development, the people that work within those structures and the priorities on which they focus all play a crucial role in determining how the evidence is considered – and these structures, people and priorities are in a constant state of flux.

Given this nature of the policymaking ecosystem, an interaction between a policymaker, a researcher and a piece of evidence rings different notes in several registers with a single strike. The intervention can connect to the project's past, present and future; the evidence resonates within the chambers of accountability of the project that preceded this one, monitoring for the project in progress and agenda-setting for the project that will ultimately succeed it. This is its iterativity. Given the flux of structures, people and priorities, this strike can also reverberate with neighbouring projects that are affected by the project under discussion – this single note ringing both in many registers and as a part of many chords. And any chord is only one stage in a larger movement, with each change of political key moving us into a new tonality. This is its complexity.

In this respect, a first major shortcoming of the bibliometric and mental reconstructive approaches is that they treat citations of research or affirmations that policy influenced a thought process as unambiguous evidence of research making a causal contribution to a specific policy outcome. The evidence collection system is asking a simple question, looking for simple answers; those answers are then used to build simple narratives, reinforcing the paradigmatic view that led to the collection system being designed in that way in the first place. These simple narratives contribute to a reductive view of policymaking, levelling off its iterative complexity (e.g., Kingdon, 2011). These measurement systems provide information that may be useful from an accountability standpoint, but this information (alone) is not easily applicable to learning purposes.

2.2: A potential solution

Ritter and Lancaster (2013) propose a measurement model that can be more useful to understand and improve the uptake of research in policy discussions. The authors acknowledge that the mechanisms for evidence uptake are iterative (and socially interactive) rather than linear (and mechanically deterministic), that these processes involve multiple stakeholder groups rather than just researchers and decision-makers and that evidence passes through multiple channels during this process rather than only from scholarly publications into policy. In terms of measurement protocol, their approach builds out from the baseline bibliometric approach.

Additionally, the authors explicitly highlight the importance of an alignment between one's conceptual understanding of how the target system functions and one's measurement system. This point is echoed and even expanded in the world of programme evaluation, where a 'logic model' or 'theory of change' is supposed to offer a point of contact between one's conceptual model of how the ecosystem functions, one's system to measure the functioning of that system and one's own interventions in that system (Pasanen and Shaxson, 2016). Endorsing these points from Ritter and Lancaster and from programme evaluation, the present report attempts to highlight and valorise these interconnections between theory, measurement and operation.

In the system they develop, Ritter and Lancaster's focus is drug policy in Australia; specifically, they seek to measure the uptake of evidence from the Illicit Drug Reporting System (IDRS) and the Ecstasy and Related Drug Reporting System (EDRS) into national drug policy. They put forward a three-component model to measure evidence uptake, each component connected to one aspect of the policymaking process as they understand it.

- **Public opinion**: Ritter and Lancaster identify public opinion as playing an important role in the policy process and furthermore that public opinion is both reflected and shaped by news media.[3] Translating that idea into a measurement protocol, they define 'media discussions' strictly in terms of newspapers, justifying their narrow focus by claiming that newspapers *"set the agenda for other news formats."*[4] In their policy context, they identified a range of key newspapers in Australia and sourced the data through Factiva.
- **Policy processes**: The authors identify discussion processes in the policy ecosystem as a key area for understanding how decisions come to be made and thus for mapping the uptake of evidence. The data source selected as a proxy for policy processes was an inventory of all public consultation submissions, as well as an inventory of all official (textual) records of discussion for federal parliamentary committee proceedings in Australia.[5]
- **Policy positions**: The authors define policy positions as the outcomes that ultimately arise from the policy process; these are the decisions at which the policy discussions ultimately arrive. As noted by Plum Analytics,[6] 'policy document' is a rather vague term and yet it must be given a clear operational definition for measurement to proceed. Ritter and Lancaster define policy documents as *"formal, publicly available iterations of position, strategy and statements of intent made by government,"*[7] and used four data sources to delineate the scope of relevant drug policy from the Australian federal government: the National Drug Strategy website, the Drug Policy Modelling Program policy timeline, the publication catalogue of the Australian Institute of Health and Welfare and the publications of the Australian Institute of Health and Welfare.

In the case of each component of their policy ecosystem model, the measurement protocol was to identify explicit mentions of the evidence from the IDRS or EDRS. As comparative benchmarks for interpreting the findings, they also tagged mentions of several other important evidence sources in their ecosystem. The search was conducted manually, with a five-year period of news coverage being analysed by two research assistants over a three-month time frame. (Resource intensiveness for policy processes and policy positions was not reported.)

2.3: Measuring up against other approaches

Compared to the baseline, the measurement approach put forward by Ritter and Lancaster does a much better job of accounting for the dynamism of policymaking. They recognise the complexity of the ecosystem and design their measurement system in consequence, broadening out the target phenomenon from policy decisions specifically to a policy discussion more generally (of which the decision is just one element and without necessarily conceptualising a decision as the end of a linear process, as decisions can spark public discourse and influence other discussion processes).

Other approaches were also discovered that similarly broadened focus from policy outcomes to policy discussions.[8] For example, we documented an analogous broadening in survey question formulation; asking respondents to rate the relevance of evidence to a given policy decision/position can be complemented by asking them to rate the relevance of the evidence to the broader policy discussion also (e.g., Bossuyt, Shaxson and Datta, 2013). Broadening the questions itself can also be complemented by broadening the population surveyed, acknowledging the wider sphere of actors who play relevant roles in the decision. Ritter and Lancaster broaden the cast of actors by including news media in their analysis.

One of the advantages of Ritter and Lancaster's suggestion, compared to approaches relying on survey or interview methods, is that the source data are created passively by ecosystem actors (as a by-product of their usual operation) rather than requiring their active participation and thereby creating an additional demand on their time.[9] That is not to say that implementing this system was without cost; as noted, the data processing was done manually. However, developments in data mining have made it such that digital systems could be built to automate the processes of harvesting source documents, extracting and parsing their content and identifying references to the evidence of interest. The cost of developing such a pipeline should not be underestimated, of course and manual data tagging may still be necessary at first (and periodically thereafter) to build/train/refine an algorithm to execute this task on an ongoing basis.

Such an automated approach is exactly what is being offered by data providers such as Altmetric and Plum Analytics. However, because the scope and scale for these providers is much larger than the focus of Ritter and Lancaster, their coverage will not be as deep in the target geographic or thematic policy area. The depth of the approach put forward here contributes considerable value from an organisational learning standpoint.

One area where Ritter and Lancaster's methodology could be improved is around the targets that they are seeking within the source documents. Admittedly, their article is not abundant in details about their search, but one gathers that they were searching for explicit mentions of the IDRS and EDRS, probably along with some variants (e.g., the name in full rather than just the acronym) and perhaps associated reports and data sets. Ritter and Lancaster note that there are a variety of pathways for evidence to find its way into policy, but there is equally well a variety of containers for the evidence to move down those pathways. For example, one might find useful traces looking for the name of an evidence-providing organisation, its employees/associates, its documents, the events it organises or in which it participates and so forth. Once again this is an area where current offerings in altmetrics will not have sufficiently deep coverage due to their focus on breadth.

Still on the topic of depth, Ritter and Lancaster focus on newspapers as a proxy for news media overall and ultimately for public opinion as well. Putting aside the question of whether or not these proxies were adequate in 2013 (when the article was published), it certainly seems that developments since then call for a richer set of proxies to be adopted now. Other forms of print media may be worth considering, along with exclusive online

content from 'official' news sources. Furthermore, social media is now a key feature of the public opinion and public discourse landscape[10] and should therefore be considered here as well. Altmetric data providers offer coverage of news and social media, demonstrating the technical feasibility of automating this process, but once again they do not offer the depth required to support in learning about the uptake of evidence into policy.

However, while newspapers and wider print sources (including their online presence) as well as social media represent important venues of public discourse, we cannot neglect the important role that is still occupied by television news – especially amongst people born before the 1980s – highlighting the importance of considering the connections between a user's perspective and the medium through which they are likely to engage.[11] In parallel, there is a proliferation of news content being provided in video and audio formats online (such as through podcasts and various online video sites, including live video feeds). Thus, while textual sources may play an important role in public discourse, other media formats are also important.

With the continuing development of voice and image recognition technology, it may at some point be possible to apply these tools to search for discussion and presentation of evidence in an even richer set of source types. With so much online content available in these formats, such analyses would continue to be relevant even after an eventual demise of television and radio – tales of which have been greatly exaggerated thus far. As noted above, this component of the approach broadens the stakeholders covered, so delineating the scope of news and social media must be done strategically in order to effectively capture the discourse of stakeholders one considers to be relevant.

One final respect in which Ritter and Lancaster's approach could be enriched would be to formulate a more detailed theory about how public discourse, policy discussion and policy decisions are interrelated and how this nexus connects with the wider evidence ecosystem. The measurement system they propose seems like a strong starting point for gathering evidence to test hypotheses. As for connection to the broader evidence ecosystem, this topic will be revisited further on in the report.

2.4: Application

In developing a measurement system that adopts the ideas put forward by Ritter and Lancaster, the following steps would likely be key building blocks:

- Who are the relevant 'publics' to consider? Where and how do these various publics converse? Where are those public conversations documented?
- Who are the relevant policymaking bodies? Which policymaking processes from these bodies are the most relevant? Where are those policy processes documented?
- Once again drawing on those relevant policymaking bodies, which policy positions from these bodies are the most relevant? Where are they documented?

Addressing these questions, for instance, Ritter and Lancaster identified a set of newspapers in Australia as their proxy (though the 'public' that this proxy was meant to measure was not explicitly clarified) and they sourced the data from Factiva. Regarding policy process and positions, Ritter and Lancaster identified federal drug policymaking bodies in Australia, along with their processes and a public data source documenting those discussions, as well as the resulting policy positions and a public data source documenting these decisions.

Similarly, the packages in which evidence finds its way into the policy sphere must be identified and a catalogue created. Ritter and Lancaster offered little detail on this point, mentioning only that they were looking for references to the IDRS and EDRS. Over and above the baseline bibliometric focus on scholarly papers, evidence transmission can happen through grey literature reports, blog posts, data sets, conference presentations or posters, interviews, events, social media activity and – notably – through online data platforms. Evidence can find its way to policy through any of these avenues and surely others and therefore each of them needs to be monitored. For example, an expert may be mentioned by name in the media, which may be an important signal of uptake.

In sum, the baseline bibliometric approach could be substantially improved by creating a wider scope of the policy ecosystem as well as the pathways for evidence to make its way into that ecosystem. This widening of scope should be guided by the operational realities, focusing on the policy ecosystems of strategic interest to them and the avenues that they themselves use to get their evidence into policy. This type of expansion puts into action the theory–measurement–operation triad endorsed by Ritter and Lancaster and encoded in programme evaluation theory.

This process of expansion can also be iterative and exploratory, setting a specific scope, examining the findings that we uncover to highlight potential new boundaries for the scope and so on. However, it is also worth considering that – given the focus on learning to improve support to policymakers – there is an indelible connection between measuring more widely and acting more widely. Put another way, in considering how widely to scope one's measurement system, one might consider whether they would take action in each of those spheres depending on the evidence they discover. If a certain policy sphere is beyond the mandate of its action, then no operational decision hinges on measurements of that sphere and it will not make operational sense to engage there regardless of what it discovers through measurement.

3. Policymaking ought to be complex

3.1: A challenge for the baseline

The concern articulated above is that the policy system does not operate in a linear fashion, challenging an assumption that is built into many measurement systems. A second concern, discussed here, is that the policy system ought not to operate in a linear fashion either. A linear conception is thus both descriptively inaccurate as well as normatively undesirable, which should inform the design of a measurement system.

This second concern is best articulated using the concept of 'issue bias,' articulated by Justin Parkhurst (2017). Issue bias occurs when discussions about evidence displace discussions about values. For example, suppose we had to decide between two programmes to fund, one (e.g., to address HIV/AIDS) about which considerable volumes of evidence exist and the other (e.g., to address Dengue fever) about which evidence is less developed. The injunction to take an 'evidence-based' approach could push us towards one programme just because more evidence is available, completely bypassing the question of which problem we consider to be most important. The debate about relative importance is rightfully a debate for society at large, not just for experts. The experts have a substantive role to play in discussing the options available, explaining the evidence and mediating discussion about that evidence. Thus, issue bias can be understood as a politically problematic use of evidence.[12]

However, when charting a course for public policy, what and how much we know about a topic should not ultimately trump how and how much we value a topic. Issue bias occurs when just such a displacement happens; it can arise in the creation of evidence (when the selection of research questions has downstream effects on shaping policy questions), in the selection of evidence (when the determination of policy priorities takes place under the guise of a discussion about how to measure outcomes) and in the interpretation of evidence (when certain methodologies are valued above others, such as randomised controlled trials being considered to the exclusion of public health data). Because all evidence embeds a perspective within it, the creation, selection and interpretation of evidence has an indelibly political component. These processes should therefore involve wider societal stakeholders, not only the experts on the evidence itself.

What does issue bias have to tell us about the impact agenda? As noted above, attribution is a central challenge of impact measurement and also a key feature of legitimating one's laying claim to the outcome. In the use of evidence for policy, however, the problem of issue bias shows us that (over and above such assumptions being descriptively inaccurate) it would be deeply undesirable to have a system where the evidence unambiguously determines the policy outcome. To believe that the evidence ought to determine the policy outcome is to embrace technocracy – overlooking the democratic need to consider, discuss and identify which evidence is relevant in a given instance. The judgement of which evidence ought to hold sway is a decision to be made by society as a whole, not by researchers (or even policymakers) alone.

3.2: A potential solution

Broadening the focus – as Ritter and Lancaster have – to encompass public discourse, policy discussions and policy decisions is a helpful start. However, even within each of these venues, it is important to valorise the different roles that evidence can play beyond a simple binary of impacted/did not impact.[13] Sandra Nutley, Isabel Walter and Huw Davies (in various publications) outline a helpful taxonomy of different uses of evidence in a policy context. In its most general formulation (Davies, Nutley and Walter, 2005), the primary distinction they draw is between instrumental use (which means driving a 'change in policy, practice or behaviour') and conceptual use (which means 'changing people's knowledge, understanding and attitudes towards social issues'). These have been diversified/complemented in specific contexts, including by other authors (e.g., Bossuyt, Shaxson and Datta, 2013; Stevens, 2011; Weiss, Murphy-Graham and Birkeland, 2005), to include variations such as the following:

- Political use, where research is used selectively, either to justify a previously concluded decision or for its fit with an evolving political narrative.
- Symbolic use, where research provides a new attitude about future priorities rather than necessarily the discussion at hand.
- Non-use, where research is relevant but difficult to take up, for conceptual, operational or political reasons.
- Misuse, where research is applied in technically incorrect ways (either intentionally or not).
- Learning use, which is a subset of instrumental use (which, by extension, suggests that accountability use might be a relevant category as well).

Here we refer collectively to these various uses under the term 'consideration' rather than 'impact.' This more neutral term seems to better reflect the normative challenge of issue bias; one might fairly presume that a given piece of evidence should be considered, even while acknowledging that in that consideration process whether it is selected and how it is carried forward are decisions legitimately in the hands of society at large.

A few approaches were identified to measure usage type (Bossuyt, Shaxson and Datta, 2013). The most common approach was interviews, where subjects were asked to identify the role that the evidence played in a given process. Another approach was the qualitative review of programme documents, with evaluators seeking to reconstruct the decision-making process and identify from the text how evidence entered into the discussion.

One automated approach from the bibliometrics world could also be imported here; in this case, Luwel and colleagues identified the in-text references to a set of papers by the section of the paper in which the citations were located, with introductory material considered primarily as conceptual use and 'non-introductory' material considered as instrumental use. However, in the case of policy documents (especially from different policymaking bodies) it may not be feasible to identify such a consistent document structure, nor to assign to document structure such a consistent and unambiguous function.[14]

One possible approach that was not encountered in this context, but seems like it might offer some promise, is the use of natural language processing. For instance, tools have been widely developed to tag text as positive, neutral or negative ('sentiment analysis'), to tag words for their grammatical function in a sentence ('parts-of-speech analysis') and to tag sentences for their rhetorical function within an overall corpus ('argument mining'). These sets of tools have taken considerable time, effort and investment from the community to be developed to their current state of maturity, so it would be worth identifying potential repurposing of these tools for the use of evidence in policy; however, evidence use in policy is likely too niche a use-case (on its own) to support the development of a new toolbox if substantial effort were required to adapt existing tools or develop new ones entirely.

3.3: Measuring up against other approaches

If we wish to valorise these various types of consideration within an operational learning context, we must find an adequate way to measure them. The concepts presented here appear to be of exceptionally high value; moving from a focus on a homogeneous type of evidence impact on policy to a heterogeneous set of consideration types enables us to map a much more nuanced landscape (and thereby also to learn a much more nuanced range of potential lessons about one's own activity in that landscape). However, while the value of these consideration types is high, the approaches to distinguish and measure them seem to be very rudimentary. Survey and interview methods come with a notable cost for both the team conducting the evaluation and the participants offering their input. Furthermore, the reliability of those reconstructions of a rationalisation process is questionable, especially when they are conducted long after the fact – a challenge that will be revisited below. Administrative documents may age better, but the manual analysis of such documents remains an expensive undertaking, while also raising questions about reliability across reviewers and review exercises.

The approach drawing on bibliometrics offers some promise in terms of the automated extraction of relevant mentions of evidence (noting that 'mentions' here should be broadened out as well in terms of the containers in which evidence find their way into policy, as discussed in relation to Ritter and Lancaster's approach). Luwel and colleagues also undertook manual sentiment analysis on the text coming just before and after the citation (extracted automatically).

3.4: Application

In order to implement such a system, it will be crucial to decide which consideration types will be adopted in the measurement system. Drawing on the theory–measurement–operation triad, this decision about which types to measure is nothing less than staking a position about the theory of consideration and putting that into practice within one's organisation. Conceptual and instrumental use are widely adopted, but as noted above there are many variations and additional types to be found spread across the literature.

One might conceive a multi-dimensional space in which each type of use could be located, which would help to provide an underlying logic to the diversity of variations. No such dimensional space was discovered and developing such a taxonomy was beyond the scope of the present project.[15] However, such a development would contribute considerable value to evolving the discussion around different types of consideration and different ways in which evidence is taken up in the policy ecosystem.

A second challenge that must be addressed, in tandem with the first but in this case more technical, is the measurement protocol for how these usage types would be identified based on excerpted text (in which the mention is found). The approaches discovered here all relied on manual coding and while natural language processing tools were highlighted as offering a promising start for developing an automated method, the present study is not in a position to anticipate what the resource requirements would be to bring text analysis tools from their current state of maturity to the point of being adequate for regular, operational use.

Strategically, however, it seems like some initial manual coding would be effort well spent. For one, this manual coding would enable quick experimentation and iteration with regard to the selection of a taxonomy of consideration types. As noted above, this choice has implications for theory, measurement and operations; accordingly, the decision itself should not be made strictly on the basis of any one of those aspects, but rather as a decision balanced on considerations coming from all three spheres.

Some manual coding of consideration types would provide valuable experience working with the raw inputs, which will help to elucidate the viability of measurement approaches, as well as the implications of this decision for theory and operational practice. Put simply, looking at real examples can help to flesh out what a certain theoretical commitment would mean in terms of the need to develop a measurement system as well as the practical implications of building a measurement system of this nature to guide operational decision-making. Lastly, manual coding provides valuable input as a training set for developing automatic coding algorithms.

4. Engaging with the evidence

4.1: A challenge for the baseline

The two main challenges presented above relate primarily to the uptake of evidence in policy, calling for measurement systems to expand: broadening the conception of the policy sphere beyond just policy decisions and broadening out from impacting those decisions to providing evidence for consideration (which then may or may not carry decisive weight in the decision-making process). However, the baseline also gives us little insight into the journey that a piece of evidence makes from production to engagement to consideration.

With regard to these 'upstream' aspects, bibliometric measures have limited information to provide, mostly citation information about the papers as well as the journals in which they appear, perhaps also some information about the age of the papers, their thematic or topical focus and their authors (and the institutions and countries in which they work). These metrics encode the considerations of the academic world, where citations and reputation within the research community constitute professional currency, the coin of the realm.

Regarding rationalisation approaches, the simplest ones will ask survey respondents or interviewees to rate the relevance of evidence to a policy decision or perhaps to the broader discussion leading up to it. In both cases, these accountability exercises are often conducted long after the fact. These approaches offer limited elucidation of the journey towards consideration (and potentially instrumental use). For accountability purposes, the impact itself is sufficient on its own. How or why the impact came about is valuable insofar as there must be a credible claim laid to the impact.

For learning, though, it is valuable to understand how and why a given piece of research was considered. Suppose that we wanted to use these bibliometric measures to derive 'actionable insights' that researchers could implement to improve their research uptake. The range of possible recommendations would extend to selecting the right journal, selecting the right topic, selecting the timing of your publication, selecting the right co-authors and so forth. In short, by measuring only the academic properties of evidence, we can articulate actionable recommendations only in academic terms.[16] However, it is widely noted that scholarly publications are not well-adapted packages through which evidence finds its way into policy (e.g., Cairney and Kwiatkowski, 2017; Cairney, Oliver and Wellstead, 2016). Accordingly, bibliometric tools to characterise those papers are on the whole going to provide unhelpful input for actionable recommendations.

Some surveys seeking to map the facilitators and barriers of uptake might ask about the features that contributed to this research being taken up, such as a plain-language summary, the timing of its publication, or the availability of an expert to discuss it. While such an approach takes a step in the right direction, what we need is a tool kit that is specifically adapted to the problem of measuring user engagement.

4.2: A potential solution

Digital marketing in the private sector has developed considerably more nuanced and sophisticated tools for measuring engagement than are commonly deployed or even discussed in the academic or public sectors. (The great irony here is that these tools are designed to enable learning; meanwhile the academic and public sectors marginalise learning in favour of accountability-first mentalities, ostensibly in an attempt to replicate the efficiency of the private sector.) Private-sector efforts to conceptualise and measure customer journeys across various touchpoints are extremely advanced and leveraging this learning offers the opportunity to take a leading role in the measurement of engagement with evidence, as an upstream signal of consideration.

Within the landscape of digital marketing, business-to-consumer (B2C) marketing is probably too transactional to be readily applicable in an evidence engagement context. However, business-to-business (B2B) marketing frameworks seem to offer considerable promise. The goals of a B2B marketer are, in many ways, similar to those of someone aiming to measure and increase the consideration of evidence. The following table illustrates

the differentiation between B2B and B2C marketing[17] in order to provide context for later parallels drawn between B2B marketing practices and evidence consideration measurement goals:

	B2C Marketing context	B2B Marketing context
Goal	**Conversion (e.g., clicking 'buy')**	**Engagement (e.g., time spent)**
Target	**Person**	**Organisation**
Method	**Convincing**	**Educating**
Journey	**Short and simple**	**Long, complex and ambiguous**
Sales cycle	**Days to weeks**	**Months to years**
Measurement focus	**Number of leads converted from each stage of the funnel to the next**	**Qualified personas exhibiting appropriate engagement signals at key touchpoints**

As we can see in the table above, B2B marketing is forced to tackle abstract and amorphous goals, ones that involve a complex system of decision-makers with various prerogatives. By contrast with B2C marketing, where optimising touchpoints for conversion (e.g., clicking the 'buy' button) is a clear and focused goal,[18] in B2B marketing, marketers have no choice but to consider the mental journeys travelled by multiple categories of users.[19] Large organisations make purchasing decisions by comparing options based on evidence that is gathered by multiple actors (and units) within the organisation and from various sources. These offer clear parallels to the case of policymaking, where evidence is over-abundant and must be marshalled to the cause of 'policy stories' as they emerge throughout the organisation and are adopted by various decision-makers (Stevens, 2011).

One consideration that should be discussed here is the extent to which individual goals are comparable from the B2B case to the evidence-for-policy case. A B2B marketer has the final goal of making a sale and thus is looking to market the evidence that they supply. Based on the ethnographic account of evidence use in government provided by Stevens, this seems not too different from the policy analyst, who wants 'their' evidence to play a key role in orienting the 'winning' narrative. In the context in which Stevens was working and in any other contexts in which his description holds, the parallel between the B2B marketer and the policy analyst seems robust.

However, no evidence suggests that his account of government culture is universal (and assumptions that a given cultural feature are universal have often turned out to be false). Furthermore, what applies to policy analysts taking up evidence does not necessarily apply to researchers who supply that evidence. Thus, the parallel between B2B marketing and the provision of evidence for policy should not be uncritically endorsed. For instance, conceptualising the policy analyst as the 'marketer' and the researcher as delivering the 'product' means that these two individuals operate in different organisations, whereas the B2B marketer and those who deliver the product or service they sell usually work in the same organisation.

These provisos now clarified, here are some concepts that form the central core of B2B marketing, as follows.

- **Persona mapping**:[20] B2B marketing often begins by clearly mapping the organisations they are targeting and the system of decision-makers inside them. This is often supported by mapping key personas – that is, creating semi-fictional representations of key decision-makers and those people that support the decision. These representations go beyond defining basic demographic aspects of the target individuals. For example, they may begin by formulating hypotheses about their position and education level, but will continue by creating realistic but fictional accounts of each persona's content consumption habits, attitudes, hobbies, fears, objections, etc. Although the final personas are meant to represent categories of users, they purposely go far beyond data that may be available on these users, in order to stimulate discussions and hypotheses that are more human-centred.
- **Customer journey mapping**:[21] Once a list of personas is clearly identified, B2B marketers build out maps of the journeys that each of these personas might take in the process of coming to their decision. The mapping covers a persona's full trajectory – not just their contacts points with the vendor – within the context of performing their role in the ecosystem. (For example, a journey may include pre-conceptions about the product/service category, media consumption habits that may have contributed to their attitude, or interactions with competitors.) Their journey is mapped individually, as well as in the context of their contribution to the broader organisational decision-making process.
- **Touchpoint analysis**:[22] Once a clear view is formulated around the journey that personas engage in, specific touchpoints – that is, points of interaction with the organisation – can be identified and analysed based on company data. Performing this analysis in the context of a persona and the broader journey they engage in enables marketers to create very targeted KPIs that not only measure engagement but measure types of engagement that are most relevant in different situations. For example, while a typical measure of engagement for a blog post might be 'time spent on page,' a touchpoint analysis based on personas and customer journeys may determine that for persona three who has gone through journey eight, engagement at this touchpoint should be measured by 'number of editing-related interactions' (e.g., copy-pasting text).

In terms of indicators, the following list covers some of the most accepted measurement categories from the B2B world.[23] They have been adapted here to an evidence-uptake context.

- **Evidence-level metrics**: Metrics that can enable organisations to gauge the level of consideration of a particular piece of evidence, such as traffic, number of downloads, scroll heatmaps, move heatmaps and click heatmaps.
- **User-level metrics**: Metrics that enable organisations to evaluate the level of engagement of specific users. Users can be uniquely identified using cookies or user profiles that require sign-in. Additionally, users can be segmented organizationally using IP tracking; for example, IP tracking was used to monitor political staffers modifying Wikipedia pages, in the United Kingdom, the United States and Canada.[24] Indicators could include time-on-site, pages per visit, number of visits or number of downloads or shares.

4.3: Measuring up against other approaches

As noted above, no other approaches encountered in the evidence-uptake realm seem to be engaged in anything very similar. The impact agenda and the focus on accountability may offer an explanation for this absence of alternatives, as the journey of evidence towards impact is only necessary insofar as it helps to bolster one's claim to impact. To the extent the linear narratives are embraced and bibliometric evidence and first-hand rationalisations are accepted as evidence of linear impact, no further evidence is called for.

Within an accountability system, forensic analyses seem like they might be the closest comparator to journey mapping, though no forensic approaches were discovered in the literature about evidence uptake (noting that no terms related to forensics were included in the initial search strategy). Ethnographic accounts, such as the one provided by Stevens, offer perhaps another angle on this topic; their treatment of political economy and all of the power brokering and other 'non-rational' behaviour that goes on in the policymaking sphere offer a completely different type of account of the journey, one that is often oriented as a critique towards linear narratives as a whole.[25]

4.4: Application

Implementing such a measurement system would require articulated personas and customer journeys. The persona, the journey and the touchpoint are all hypotheses and they must be tested with respect to outcomes later in the process – namely, in this case, the consideration of evidence. In this way, the mapping of personas, journeys and touchpoints is dynamic rather than static. In the B2B context, this mapping is also dynamic, though in that case the trend is primarily for new personas and journeys to emerge – creating additional maps rather than revising existing ones.

5.2: A potential solution

What is proposed here is an integration of two approaches encountered during the literature review: the SEER tool and the theory of planned behaviour. Each of these will here be presented in turn, followed by a proposed integration of them.

5.2.1 Seeking, Engaging with and Evaluating Research (SEER) tool

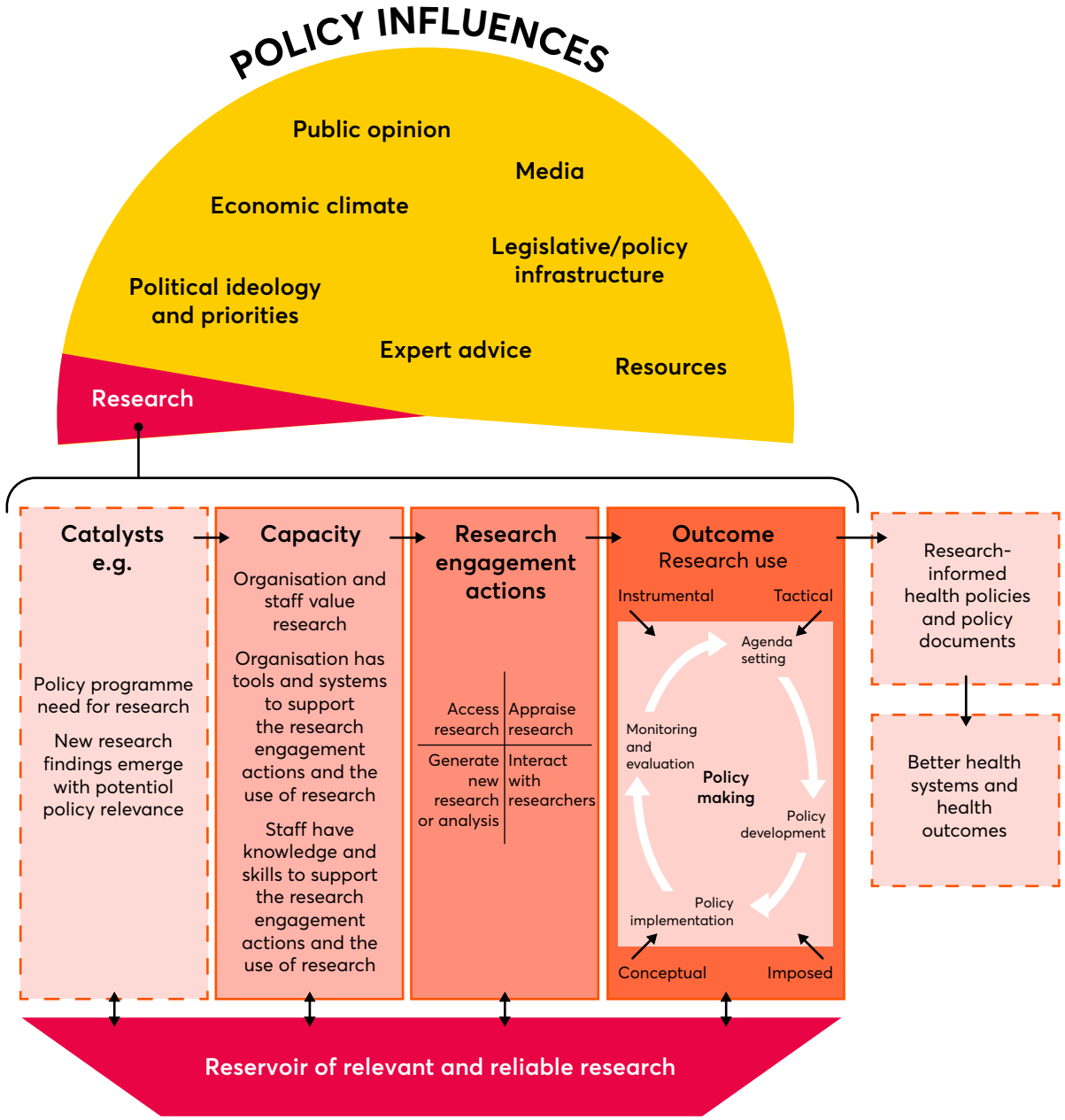
The Seeking, Engaging with and Evaluation Research (SEER) tool is a self-report questionnaire designed to assess individual policymakers'

1. Capacity to use research (predisposing factors),
2. Research engagement actions (research was accessed, appraised, generated, etc.) and
3. Research use (research was used instrumentally, conceptually, etc.).

It consists of 50 questions scored on either Likert or binary (yes/no) scales. Aggregate scores on research capacity are hypothesised to predict engagement actions and use. The SEER tool underwent a validation procedure involving 150 participants engaged in policymaking (Brennan et al., 2017). The scales showed acceptable internal validity and reliability and were able to predict some proxies of research use, although the authors suggested that further work needed to be done on refining the measures of research engagement and use. The SEER tool thus appears to be a promising measure of capacity to engage with research.

The tool was developed by Brennan et al. (2017) and was designed to measure the impact of capacity-building strategies to increase the use of research in health policy development. It was framed within the SPIRIT (Supporting Policy In health with Research: an Intervention Trial) Action Framework, which is an Australian initiative to develop and test interventions that support research use in policy agencies (Redman et al., 2015). The SPIRIT Action Framework identifies three main factors related to research uptake in policymaking: capacity, engagement and use. The SPIRIT framework thus aligns well with the synthesis of the relevant literature on research uptake measurement offered in this report, including elements such as a non-linear conceptualisation of the research-to-policy process and a diversity of consideration types.

Figure 1. The SPIRIT Action Framework



Source: Reproduced from Redman et al. (2015).

Two other tools besides SEER have been developed within the SPIRIT framework: ORACLe (Organisational Research Access, Culture and Leadership) and SAGE (Staff Assessment of enGagement with Evidence; Makkar et al., 2015; R. Makkar et al., 2016). The ORACLe tool consists of a structured interview of 23 questions given to leaders of organisations engaging in research use and policymaking. The interviews are recorded, transcribed and scored

according to a standardised scoring system by third-party coders who were not directly involved in the interview process. The SAGE tool consists of an analysis of a policy or programme document produced in the last six months and a semi-structured interview of 22 questions with a policymaker who contributed significantly to the document's development. The interviews are transcribed and scored along with the policy document itself using a standardised scoring system, ideally performed by a third party.

One positive aspect of all of these tools is that they are domain-general and could be used for contexts outside of health research. The ORACLe, SAGE and SEER tools are complementary in that they target research capacity at different levels, from organisational leaders (ORACLe), to those directly engaged in recent policy-document writing (SAGE), to staff engaged in policymaking in general (SEER). While all three tools are valuable in their own right, the ORACLe and SAGE tools require a more intensive research process of interviews, transcriptions and scoring by third parties, whereas the SEER tool is an entirely self-reported questionnaire that can be implemented automatically within an online platform (Brennan et al., 2017). The SEER tool thus appears to be a more feasible tool to implement, although targeted use of the ORACLe or SAGE tools may still be informative depending on research capacity and goals.

5.2.2 Theory of Planned Behaviour

The theory of planned behaviour (TPB) was developed by Boyko, Lavis, Dobbins and Souza (2011). Whereas the SEER tool aims to measure capacity to use research (an organisational focus), TPB aims to measure intention to use research (a psychological focus). The TPB tool is a 15-item self-report questionnaire that measures factors hypothesised to predict research use in health policymaking. The factors are based on concepts related to TPB:

1. **Attitudes** (e.g., *"I want to use research."*),
2. **Subjective norms** (e.g., *"People I care about want me to use research."*),
3. **Perceived behavioural control** (e.g., *"I have the power to use research."*) and
4. **Behavioural intentions** (e.g., *"I intend to use research."*).

The first three factors are hypothesised to predict the fourth, which in turn predicts actual behaviour. The TPB tool has been shown to predict behaviour in a number of contexts, including the use of research evidence in healthcare professionals (Boyko et al., 2011).

The psychometric properties of the TPB tool have not been studied extensively, as the original study introducing the tool suffered from a small sample size and did not include outcome variables that measured actual behaviour (Boyko et al., 2011). However, the study that implemented the SEER tool also asked participants to complete the TPB tool (Brennan et al., 2017). The results showed that the components of the TPB tool predictive of research

use (i.e., behavioural intention to use research) and the TPB questions that measure similar constructs as the SEER tool (e.g., attitudes towards research) correlated with the relevant portions of the SEER capacity scales. It also predicted indicators of research engagement actions and research use (in many cases to a greater degree than the SEER capacity scales). This indicates that the TPB tool shows promise in being able to predict behaviours related to research engagement and use, although an analysis of the TPB tool's internal validity would be recommended prior to implementing it in a new context, as discussed below.

5.2.3 Combining SEER and TPB

The challenge here is to overcome the gap between engagement with evidence and the downstream consideration of that evidence. In terms of measurement, there is not only a loss of signal between platform engagement and downstream measures, but the two are also separated by a lag in time. If engagement with evidence through online platforms were to increase substantially, the downstream changes might take months or years to appear and it would be difficult to relate an uptick in consideration months down the line to a specific change in the platform that may have prompted it (if it was even a change to the platform rather than a change elsewhere in the policy ecosystem that prompted this uptick).

The SEER and TPB tools were designed with the aim of measuring the before/after impacts of interventions to support evidence-based policymaking. Elements from these tools could be combined into a single short survey (e.g., about five questions) that users would be prompted to answer during their use of an online platform. Such an approach would mirror user-experience (UX) questionnaires, which are designed to assess how the user interacts with a site, while only requesting a small effort on their part. Given the value already demonstrated for the SEER and TPB tools, such an approach seems to offer the potential of a leading indicator of evidence consideration, in effect offering a bridge between the upstream and downstream components of the measurement system articulated thus far.

If shorter versions of the tools were to be made, it would be best to take a statistically informed approach, building on the items that appear to be the most valid indicators of the constructs measured. Brennan et al. (2017) covered the psychometric properties of SEER, but not those of the TPB tool. Creating shorter tools with high internal validity could be done using a factor-analysis approach, where the most pertinent items to use would be determined by the strength of association between each original item to the underlying construct (Swisher, Beckstead and Bebeau, 2004). Items with the strongest association with the underlying construct would most likely be the best to use in the shorter versions. Given that the internal factor structure of the TPB tool has not been published, this would require analysing the data collected by Brennan et al (2017), which would need to be requested (as searches to date have failed to locate a data set available online). Creating such a tool could be valuable for the policy research community as a whole; these tools are sought after, as attested by the SPIRIT framework studies.

5.3: Measuring up against other approaches

Unlike the baseline, the approach outlined here would offer tangible clues about the pathway from engagement with evidence to its eventual consideration in a policy setting. As noted above, the most routine bibliometric approaches offer no clarity on the path traversed between evidence publication in a scholarly journal and evidence citation in a policy document. Rationalisation approaches mainly consist of asking people, either in a survey or an interview, to reconstruct a given process of reasoning and assess the role of evidence in that process – usually well after it has taken place.

However, there are good reasons to wonder about what exactly goes into those retrospective reconstructions of a rationalisation process. As Kahneman (2013) argues (along with many others in behavioural science), subjective confidence in memory is connected primarily to the felt coherence of that memory – both its internal coherence among the pieces of information one retrieves as well as its coherence with a wider context of one's memories and sense of self. Confidence in a given memory is not a reliable indication that it presents a good quantity or quality of information. In fact, it is easier to build a coherent picture out of less information than out of more information, because there are fewer potential discords that might need to be resolved. Furthermore, framing effects and question substitution (as well as other biases and heuristics) can occupy a larger space in shaping one's thought process when only little information is available in memory to fill that space.

What does all this mean in the present context? Both the baseline approach and the approach suggested here using SEER and TPB rely on an assumption: that you can get valuable data by asking someone to evaluate the role that a given piece of information played in their own thought process. While there may be concerns about this assumption (concerns that apply even to the approach proposed here), there are additional concerns about asking the question long after that thought process has taken place. Our recollection of information fades over time, which increases our susceptibility to biases and heuristics but without decreasing our subjective confidence in the accuracy of what we recall. The proposed approach therefore at least offers a better chance at collecting valuable data, even acknowledging the very real concerns about how transparently we understand our own decision-making even in the first person.[26]

Concerns of this kind provide some of the motivation for a move towards ecological momentary assessment (EMA), which *"assesses individuals' current experiences, behaviors and moods as they occur in real time and in their natural environment"* (Burke et al., 2017). As discussed above, the focus of EMA on real-time measurement aligns with the approach suggested here. The measurement within the 'natural' environment also aligns with what is proposed here, as the short survey would be administered to users during their engagement with an evidence platform, which is the environment of highlighted interest here. One of the benefits of EMA therefore, one that is also applicable in the present case, is that it creates the possibility to link users' feedback to their actions within the environment just before and after their response.

Basically, if evidence providers instrument their platforms (as described in the previous section) to measure user engagement and implement a short user survey (as described here) to assess intention to consider evidence, then cross-linking these data sets would enable them to map out how various user behaviours contribute to the consideration to use evidence. Such knowledge seems invaluable in refining the platform to optimise consideration. Here it is worth once again noting the ethical ambiguities at play in this situation; setting up these types of measurement systems can prompt strong goal-seeking behaviour on the part of data scientists and platform designers, underscoring the importance of establishing an appropriate goal from the outset.

5.4: Application

As noted above, a key component of implementing these uptake measurement tools would be to develop a short version of the SEER and TPB tools that could be implemented as a user survey on data platforms. Such a survey could supply useful key performance indicators (KPIs) as a compass to guide A/B testing of platform design options. Eventual downstream actions (i.e., the consideration of evidence in policy) could also be traced back to scores on these tools to double-check their predictive validity.

Either instead of or in parallel with the deployment of a shortened survey, applying the full-length tools with a selection of users might be worthwhile; engaging in targeted semi-qualitative surveys of key users can be a valuable learning exercise in its own right and has the potential to build bridges between evidence producers and users. Previous attempts at increasing research uptake in a Canadian healthcare context using semi-structured interviews with key stakeholders did not show much predictive validity in subsequent research uptake (Dobbins et al., 2009; Kothari, Edwards, Hamel and Judd, 2009). Nevertheless, the exercise of discussing research uptake capacity using the semi-quantitative survey was seen as helpful in starting valuable discussions about research-based policymaking within the organisations studied.

Semi-structured interviews also have the potential to uncover important insights that might be missed by an overly constrained quantitative tool. Quantitative and qualitative approaches each have their strengths and weaknesses, so a trade-off is always made in selecting a given tool.

6. Conclusion

6.1: Four challenges and components for a measurement system

The present report identifies two approaches for measuring the impact of evidence on policy: using bibliometric tools to identify citations to scholarly papers within policy documents and using survey or interview techniques to elicit judgements from stakeholders about the role of evidence in a process of decision-making. Four major lines of critique are outlined – that a linear model of policymaking is neither (§2) descriptively accurate nor (§3) normatively desirable, (§4) that these measurement approaches offer very little insight into how policymaker users are actually engaging with research and (§5) that they similarly tell us very little about the journey from engagement to consideration.

In response, this report articulates several component parts that could be piloted in developing a measurement system. In response to the challenges to the linear model of policymaking, the suggestions put forward are to examine a wider range of source types, expanding well beyond scholarly journal articles to include many written forms as well as non-textual sources (§2). This suggestion acknowledges that evidence comes in many packages. Furthermore, signs of uptake need to be captured in a much wider range of source documents, reflecting the diversity of actors and discussion venues involved in policymaking. Also, how the evidence is used needs to be considered in a more nuanced light, moving beyond the binary of impact/no impact (so easily mapped onto the binary of cited/not cited) and acknowledging a fulsome breadth of types for evidence consideration (§3).

Regarding the measurements further 'upstream' in the process, the report puts forward suggestions for measuring user behaviour on a self-service platform for evidence exploration (§4); these measures will help in understanding how users are engaging with the evidence (and some increased sharing functionalities might help those users to also broadcast their policy narratives more widely and more effectively). To connect platform engagement with 'downstream' consideration of evidence in the policy sphere, the report highlights some very promising possibilities to adapt existing long-form questionnaires into condensed tools that can be implemented on the platform as short surveys, which – given a connection to actual evidence consideration – can offer a leading indicator for assessing platform performance and arbitrating between candidate revisions to the platform design (§5).

7. Methodology

7.1: Initial literature query

To identify the major trends in measuring evidence uptake, The Decision Lab undertook a review of scholarly work as well as 'grey' literature (such as reports and blogs, from the public, private and academic sectors). Given that peer-reviewed literature is more systematised, approaches for discovery and prioritisation of this literature were more methodical. Our initial searches were conducted using combinations of keywords such as the following:

Evidence	**Uptake**	**Evaluation**
Research	**Usage**	**Policy**
Measurement	**Impact**	**Innovation**

Several hundred results were given a manual 'first-pass' inspection for relevance, based on the titles and keywords. This preliminary investigation supplied additional keywords to test in our queries, as well as exclusion criteria when large bodies of literature were discovered that were beyond the scope of this project (such as measuring the impact of a policy on target stakeholders, which was a major neighbouring topic, but further downstream than the focus of this mandate).

Based on these inspections, an initial list of about 120 papers was established, with the majority published during the 2015–2019 period. In addition to assessments of research uptake in general (agnostic as to topic area), there were two clear thematic areas of focus: one on health research and one on international development/aid research (influencing their respective areas of policy). Papers on this list had many authors in the United Kingdom, the United States, Australia and Canada; even papers focusing on international development/aid usually included authors from the industrially developed anglosphere, though uneven digital coverage of journals by country of publication, language and thematic area may have contributed to this pattern.

To validate and expand the scope of this initial list, an automated approach to research discovery was applied. The 'Scholar's reading list' tool[27] enables a user to identify and characterise clusters within a body of research. However, the tool is designed to be used on much larger bodies of literature, aiming at an order of magnitude of ~10,000 papers rather than ~100. Accordingly (with the generous help of Maxime Rivest, the creator of the Scholar's reading list), the initial 'core' list of papers was expanded to include all those papers cited by an article on the list, all those papers citing an article on the list and all those articles that cite the same papers as those on the list.

Once this expansion was undertaken, the complete list included several thousand papers. By applying the clustering and characterisation tools, we identified several distinct topics within the landscape of this literature. Several of these topics were beyond the scope of relevance, such as papers assessing the extent to which a change in health policy impacted

patient outcomes. Other topics identified were highly relevant and most of these were already well covered in our core list, validating the initial literature discovery strategy.

One novel topic emerged that was on the boundary of the project scope: knowledge management. Much of this literature focuses on internal dynamics within individual organisations, considering the incentive and management structures that an organisation can put in place to increase the flow of knowledge across divisional boundaries (and over time as employees shift to new teams). The knowledge management literature was therefore reviewed for this project, but not in the same depth as the other topics. This literature holds promise, however, to offer valuable inspiration for the broader discussions here, even though it is not directly applicable.

Works cited

American Society for Cell Biology. (2013). *San Francisco Declaration on Research Assessment (SF DORA)*. Retrieved from: https://sfdora.org

Bossuyt, J., Shaxson, L. and Datta, A. (2013). *Study on the uptake of learning from EuropeAid's strategic evaluations into development policy and practice*. Retrieved from European Commission website: https://ec.europa.eu/europeaid/sites/devco/files/uptake-study-main-report-2013-317962_en_.pdf

Boyko, J. A., Lavis, J. N., Dobbins, M. and Souza, N. M. (2011). Reliability of a tool for measuring theory of planned behaviour constructs for use in evaluating research use in policymaking. *Health Research Policy and Systems, 9*(1). https://doi.org/10.1186/1478-4505-9-29

Brennan, S. E., McKenzie, J. E., Turner, T., Redman, S., Makkar, S., Williamson, A., … Green, S. E. (2017). Development and validation of SEER (Seeking, Engaging with and Evaluating Research): a measure of policymakers' capacity to engage with and use research. *Health Research Policy and Systems, 15*(1). https://doi.org/10.1186/s12961-016-0162-8

Burke, L. E., Shiffman, S., Music, E., Styn, M. A., Kriska, A., Smailagic, A., … Rathbun, S. L. (2017). Ecological momentary assessment in behavioral research: addressing technological and human participant challenges. *Journal of Medical Internet Research, 19*(3), e77. https://doi.org/10.2196/jmir.7138

Cairney, P. and Kwiatkowski, R. (2017). How to communicate effectively with policymakers: combine insights from psychology and policy studies. *Palgrave Communications, 3*(1). https://doi.org/10.1057/s41599-017-0046-8

Cairney, P., Oliver, K. and Wellstead, A. (2016). To bridge the divide between evidence and policy: reduce ambiguity as much as uncertainty. *Public Administration Review, 76*(3), 399–402. https://doi.org/10.1111/puar.12555

Campbell, D., Tippett, C., Struck, B., Lefebvre, C., Côté, G., St-Louis Lalonda, B., … Archambault, É. (2017). *Knowledge and technology flows in priority domains within the private sector and between the public and private sectors*. Retrieved from https://publications.europa.eu/en/publication-detail/-/publication/44501b5a-77f6-11e7-b2f2-01aa75ed71a1/language-en/format-PDF

Carden, F. (2009). *Knowledge to policy making the most of development research*. Los Angeles: Sage.

Davies, H. T. O., Nutley, S. and Walter, I. (2005). *Assessing the impact of social science research: conceptual, methodological and practical issues*. Retrieved from Research Unit for Research Utilisation website: https://www.odi.org/sites/odi.org.uk/files/odi-assets/events-documents/4381.pdf

Garfield, E. (1979). *Citation indexing, its theory and application in science, technology and humanities. In A Wiley-Interscience Publication*. New York: Wiley.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S. and Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429–431. https://doi.org/10.1038/520429a

Kahneman, D. (2013). *Thinking, fast and slow*. New York : Farrar, Straus And Giroux. Print.

Kingdon, J. W. (2011). *Agendas, alternatives and public policies* (Updated 2nd ed). In *Longman Classics in Political Science*. Boston.

Makkar, S. R., Turner, T., Williamson, A., Louviere, J., Redman, S., Haynes, A., … Brennan, S. (2015). The development of ORACLe: a measure of an organisation's capacity to engage in evidence-informed health policy. *Health Research Policy and Systems, 14*(1). https://doi.org/10.1186/s12961-015-0069-9

Makkar, S., Brennan, S. R., Turner, T., Williamson, A., Redman, S. and Green, S. (2016). The development of SAGE: A tool to evaluate how policymakers' engage with and use research in health policymaking. *Research Evaluation, 25*(3), 315–328. https://doi.org/10.1093/reseval/rvv044

Martin, S., Nutley, S., Downe, J. and Grace, C. (2016). Analysing performance assessment in public services: how useful is the concept of a performance regime? *Public Administration, 94*(1), 129–145. https://doi.org/10.1111/padm.12206

Mellon, J. and Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics, 4*(3), 205316801772000. https://doi.org/10.1177/2053168017720008

Parkhurst, J. O. (2017). *The politics of evidence: from evidence-based policy to the good governance of evidence*. London; New York: Routledge.

Pasanen, T. and Shaxson, L. (2016). *How to design a monitoring and evaluation framework for a policy research project*. Retrieved from Overseas Development Institute website: https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/10259.pdf

Redman, S., Turner, T., Davies, H., Williamson, A., Haynes, A., Brennan, S., … Green, S. (2015). The SPIRIT Action Framework: A structured approach to selecting and testing strategies to increase the use of research in policy. *Social Science & Medicine, 136–137*, 147–155. https://doi.org/10.1016/j.socscimed.2015.05.009

Ritter, A. and Lancaster, K. (2013). Measuring research influence on drug policy: A case example of two epidemiological monitoring systems. *International Journal of Drug Policy, 24(1)*, 30–37. https://doi.org/10.1016/j.drugpo.2012.02.005

Schot, J. and Steinmueller, W. E. (2018). Three frames for innovation policy: R&D, systems of innovation and transformative change. *Research Policy, 47*(9), 1554–1567. https://doi.org/10.1016/j.respol.2018.08.011

Shaxson, L. (2019). Uncovering the practices of evidence-informed policy-making. *Public Money & Management, 39*(1), 46–55. https://doi.org/10.1080/09540962.2019.1537705

Stevens, A. (2011). Telling policy stories: an ethnographic study of the use of evidence in policy-making in the UK. *Journal of Social Policy, 40*(2), 237–255. https://doi.org/10.1017/S0047279410000723

Sutherland, W. J., Goulson, D., Potts, S. G. and Dicks, L. V. (2011). Quantifying the impact and relevance of scientific research. *PLoS ONE, 6*(11), e27537. https://doi.org/10.1371/journal.pone.0027537

Swisher, L. L., Beckstead, J. W. and Bebeau, M. J. (2004). Factor analysis as a tool for survey analysis using a professional role orientation inventory as an example. *Physical Therapy, 84*(9), 784–799. https://doi.org/10.1093/ptj/84.9.784

Weiss, C. H., Murphy-Graham, E. and Birkeland, S. (2005). An alternate route to policy influence: how evaluations affect D.A.R.E. *American Journal of Evaluation, 26*(1), 12–30. https://doi.org/10.1177/1098214004273337

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., … Johnson, B. (2015). *The metric tide: report of the independent review of the role of metrics in research assessment and management*. https://doi.org/10.13140/rg.2.1.4929.1363

Wimpenny, P., Johnson, N., Walter, I. and Wilkinson, J. E. (2008). Tracing and identifying the impact of evidence – use of a modified pipeline model. *Worldviews on Evidence-Based Nursing, 5*(1), 3–12. https://doi.org/10.1111/j.1741-6787.2007.00109.x

Endnotes

1. https://www.digital-science.com/blog/uncategorised/360000-policy-documents-integrated-into-dimensions
2. https://plumanalytics.com/plumx-now-includes-policy-document-citations
3. They do not, however, explore in much detail how public opinion influences the policy process; this is a point that may be worth exploring further when developing measurement systems in contexts where public opinion is believed to be especially influential on the policy process.
4. While the authors do reference research that supports this position, that research is from the early 2000s. The evolution of the news landscape since that time surely calls for a reassessment of this assumption to see whether it still holds and accordingly whether newspapers alone can offer an adequate sample of the overall signal. In particular, it is highlighted that measuring social media uptake might be an important component to integrate now as well, though the question of whether to treat traditional and social media separately is left open here.
5. https://parlinfo.aph.gov.au/parlInfo/search/search.w3p
6. https://plumanalytics.com/plumx-now-includes-policy-document-citations
7. It is also important to acknowledge how much policy discussion takes place beyond the boundary of what is publicly available, even things so mundane as the early versions of documents that eventually become public when they mature. Such private communications, while highly informative as evidence to understand policymaking and the role of evidence, are beyond the scope of probably any measurement approach. These meetings, workshops and other processes may provide some proxy for the private communication, but ultimately those exchanges remain private.
8. On a closely related note, Sutherland and colleagues (Sutherland, Goulson, Potts and Dicks, 2011) started from a set of policy prescription, identified a body of literature that touched on this topic and then asked a panel of experts to score each paper based on the strength of the evidence it offered and the fit with the policy prescription. This approach, promoted by Cambridge University (https://www.publicpolicy.cam.ac.uk/pdf/policy-impact-april-2017), actually measures the potential for evidence to be taken up rather than its involvement in the policy discussion whatsoever and only relative to a problem that is clearly defined from the outset.
9. In many instances, a key barrier for programme evaluations to support learning was that operational programme staff were not sufficiently involved in the evaluation process. On its face, such a point may seem at odds with the valorisation of passive approaches, which seek to minimise burden rather than to maximise involvement. However, lack of operational staff participation was usually cited as a challenge because the people who were involved were primarily involved in budget decisions rather than operations and accordingly this differential participation allowed the accountability focus to displace a focus on learning. In such cases, operational staff participation is often perceived as an administrative burden (Martin, Nutley, Downe and Grace, 2016) rather than a meaningful opportunity. Thus, 'passive' systems will be preferred here, noting that when opportunities are available to learn and put that learning into practice, participation in these measurement exercises can offer tangible benefit. And in any case, active participation should be scoped strategically, engaging stakeholders in the process at the point where their involvement provides the greatest benefit (both locally and globally within the ecosystem).

10. While the importance of social media should not be underestimated, it should also not be overestimated; these platforms are important fora for discourse, but they are not representative samples of the population or therefore of public opinion at large (Mellon and Prosser, 2017).
11. http://www.digitalnewsreport.org
12. By contrast, 'technical bias' refers to a scientifically problematic use of evidence, whether intentional or not. For instance, constructing a sample without controlling for a potential source of systemic error would be a technical bias not an issue bias.
13. Ritter and Lancaster acknowledge this point; e.g., *"mentions of research within documents, processes or media do not mean that the research changed policy: the logic model is that in the first instance, research needs to be noticed and engaged with prior to it being able to have impact."* They draw on Nutley's work, just as we do here. However, it was worthwhile to go back to the original source, both because Ritter and Lancaster's presentation of this topic is rather undetailed and because Nutley's influence has spread much wider than Ritter and Lancaster alone.
14. Along similar lines, one might conceive of this conceptual taxonomy embedded not within the structure of individual documents, but across document types.
15. While by no means exhaustive, a list of dimensions could include instrumental vs. conceptual use, appropriate vs. inappropriate use (considering both issue bias and technical bias, noting that Parkhurst also outlines three stages at which each type of bias can arise), use for learning vs. accountability purposes and even perhaps use vs. non-use.
16. This challenge echoes a wider concern: much of the analysis of evidence uptake is written from the perspective researchers, with researchers often playing the role of protagonist in the impact narratives they create. This challenge motivates a critical reading of the discourse, of the type provided here and valorises ethnographic approaches to explaining evidence uptake, as these approaches aim to prioritise the voice of participants rather than the ethnographer. (Shaxson, 2019)
17. https://sujanpatel.com/content-marketing/b2b-vs-b2c
18. It is acknowledged here that even in B2C, there are complexities related to customer lifetime value; some businesses seek to maximize customer returns to the site and repeat purchasing, whereas others focus more on optimizing cart size within a single visit because they anticipate only few repeat purchases. In either case, though, all of the desired actions pass through the key touchpoint of the 'buy' button.
19. https://neilpatel.com/blog/b2b-content-strategy
20. http://jem9.com/b2b-customer-personas-template
21. https://customerthink.com/14-visualizations-mapping-the-b2b-buyer-journey; https://econsultancy.com/how-to-use-econsultancy-s-b2b-customer-journey-mapping-canvas
22. https://medium.com/the-marketing-playbook/understanding-customer-experience-in-saas-a9d7550c157e
23. https://www.snapapp.com/blog/b2b-marketing-metrics; https://www.leadfeeder.com/blog/b2b-google-analytics-guide; https://www.seerinteractive.com/blog/how-hotjar-works
24. https://www.theguardian.com/technology/2014/jul/30/how-to-find-out-when-uk-politician-edits-wikipedia-page; https://en.wikipedia.org/wiki/CongressEdits; https://www.cbc.ca/news/politics/political-staffers-best-be-wary-when-wrangling-wikipedia-entries-1.2721670
25. Continuous with its focus on behavioural science, The Decision Lab endorses neither a totally rational nor a totally irrational picture of decision-making. Rather, rational and non-rational elements of decision-making are both present in the ecosystem and it is an empirical question about which mixture of features contributes to a given outcome.
26. Asking a user about their intention to use evidence might act as a primer – in which case the measurement itself could be a predictor of evidence use because it is a partial cause of evidence use.
27. http://scholarsreadinglist.com

About Nesta

Nesta is an innovation foundation.

For us, innovation means turning bold ideas into reality and changing lives for the better. We use our expertise, skills and funding in areas where there are big challenges facing society.

Nesta is based in the UK and supported by a financial endowment. We work with partners around the globe to bring bold ideas to life to change the world for good.

www.nesta.org.uk

About The Decision Lab

The Decision Lab is a Canadian think-tank dedicated to democratising behavioral science through research and analysis. They apply behavioural science to create social good in the public and private sectors.

www.thedecisionlab.com

If you'd like this publication in an alternative format such as Braille, large print, please contact us at: information@nesta.org.uk

58 Victoria Embankment
London EC4Y 0DS

+44 (0)20 7438 2500
information@nesta.org.uk
@nesta_uk
www.facebook.com/nesta.uk
www.nesta.org.uk