Article

Evaluation
2021, Vol. 27(1) 116–137

DOI: 10.1177/1356389020978490
journals.sagepub.com/home/evi
SAGE

Cased-based modelling and scenario simulation for ex-post evaluation

Corey Schimpf
University at Buffalo – The State University of New York, USA

Pete Barbrook-Johnson
University of Surrey, UK

Brian Castellani
Durham University, UK

Abstract

Despite 20 years of increasing acceptance, implementing complexity-appropriate methods for ex-post evaluation remains a challenge: instead of focusing on complex interventions, methods need to help evaluators better explore how policies (no matter how simple) take place in real-world, open, dynamic systems where many intertwined factors about the cases being targeted affect outcomes in numerous ways. To assist in this advance, we developed *case-based scenario simulation*, a new visually intuitive evaluation tool grounded in a data-driven, case-based, computational modelling approach, which evaluators can use to explore counterfactuals, status-quo trends, and what-if scenarios for some potential set of real or imagined interventions. To demonstrate the value and versatility of case-based scenario simulation we explore four published evaluations that differ in design (cross sectional, longitudinal, and experimental) and purpose (learning or accountability), and present a prospective view of how case-based scenario simulation could support and enhance evaluators' efforts in these complex contexts.

Keywords

case-based methods, computational social science, evaluation, policy, scenario analysis, social complexity

Corresponding author:
Corey Schimpf, Department of Engineering Education, University at Buffalo – The State University of New York, 140 Capen Hall, Buffalo, NY 14260-5030, USA.
Email: schimpf2@buffalo.edu

Introduction: Policy evaluation 'in' a complex world

Despite 20 years of increasing acceptance and widening usage, implementing a complexity-appropriate approach to ex-post policy evaluation remains, at a practical level, a significant challenge. As Moore et al. (2019) outline, practical issues range from how, exactly, one decides what policies should be evaluated as complex – or, alternatively, which evaluations should be seen through a 'complex systems theory' lens – to how to commission and manage such 'complexity-appropriate' evaluations.

The other big issue, which holds our focus here, is method – this includes how best to think about evaluation from a complex systems perspective, as well as deciding, in turn, which 'complexity-appropriate' evaluation designs and methods one should use. More specifically, it means realizing that conceptualizing evaluation in complex systems terms does not necessarily lead to modelling only complex interventions. Instead, as Moore et al. (2019) outline, it often also means realizing that policies (no matter how simple) take place in complex systems. For us, this practical difference in conceptualization, while subtle, leads to a very real difference in how one uses the available repertoire of complexity-appropriate methods.

We articulate this difference as follows. Instead of only developing systems approaches to ex-post evaluation, we agree with Moore et al. (2019) and others (e.g. Gilbert et al., 2018) that current methods also need to better help evaluators explore, in an accessible and reasonably simple manner, how policies take place in real-world open and dynamic systems, where (1) many intertwined factors about the cases being targeted affect outcomes in numerous ways, (2) many insights may only be available retrospectively and (3) controlling or eliminating all unpredictable events is unattainable (Schimpf and Castellani, 2020). Hence, the purpose of the current paper. We have been working, over a series of papers, to develop a new mixed-methods approach for *thinking in complex systems terms about interventions in complex systems* (Barbrook-Johnson et al., 2019; Castellani et al., 2019; Schimpf and Castellani, 2020). This approach is based, in part, on our integration of two key methodologies: scenario analysis (Morell, 2005; Schwartz, 1991) and case-based methods (CBMs; Byrne and Ragin, 2009), specifically case-based computational modelling. We call this integrated approach *case-based scenario simulation* (CBSS). Scenario analysis is widely used in evaluation research (Gates, 2016; Love and Russon, 2000; Morell, 2005), while the latter is gaining prominence (Befani et al., 2007; Pattyn et al., 2019; Verweij and Gerrits, 2013). Case-based computational modelling is particularly useful because it leverages advances in computational social science (e.g. machine learning, cluster analysis, data visualization, data forecasting) to engage in a case-based approach to policy evaluation (Barbrook-Johnson et al., 2019; Castellani et al., 2019; Schimpf and Castellani, 2020).

The web-based platform for CBSS is COMPLEX-IT, which is designed to provide evaluators quick, flexible access to a range of computational modelling methods to explore their evaluation from a CBSS viewpoint (Schimpf and Castellani, 2020). We believe this approach holds potential value for evaluator as it can: (1) provide a broader landscape of evaluation methods, design, and questions that are highly synergistic with other complexity-appropriate and traditional evaluation methods; (2) offer a means to iteratively explore and interrogate evaluation data through a case-based perspective from several angles of inquiry; and (3) align with existing evaluation approaches (most strongly, realist evaluation) to examine 'what

works for whom in which circumstances' (Pawson and Tilley, 1997: 77), by modelling unique case types, their trajectories and the heterogeneous effects interventions may have on them.

To demonstrate the potential of this approach to the widest audience possible, we decided that, rather than apply CBSS to just one case study, we would demonstrate how it could be used to enhance several different policy scenarios. As such, we chose four already published evaluations that vary in design (cross sectional, longitudinal, and experimental) and purpose (learning or accountability). The trade-off in this approach is that we cannot provide an exhaustive review. However, the advantage is that the utility of CBSS can be more widely appreciated. The paper is structured as follows. We begin by reviewing the evaluation literature in regard to CBM and scenario analysis. We then present a quick introduction to CBSS and its web-platform, COMPLEX-IT. We then review the four different evaluation studies, showing how CBSS could be used to assist their evaluation. We end with a discussion of CBSS's potential as an evaluation tool and the broader implications for using case-based methods in evaluation studies.

Literature review: Modelling cases and simulating scenarios

CBMs are increasingly deployed in policy and programme evaluation studies (Thiem, 2014, 2017). Perhaps the most established and widely used approach is case studies (Bowering, 1984; Garaway, 1996; Hayton, 2015; Leone, 2008; Van Draanen, 2016; Vellema et al., 2013; Yin, 1997, 2013). Case studies predate many other CBMs and thus exhibit several properties that were adapted into more recent case-comparative methods. For all case studies, these properties include holistic analysis of cases as distinct entities (Yin, 1997) and resonance with realist evaluation (Koenig, 2009) and for multiple case studies an emphasis on case variation, including similarities and differences in case profiles and trajectories (Cousins and Bourgeois, 2014; Mookherji and LaFond, 2013; Savaya et al., 2008).

Another increasingly prominent technique, qualitative comparative analysis (QCA), employs a set-theoretic approach to identify combinations of causal conditions from cases profiles (i.e. variables) that relate to outcomes of interest. QCA has received considerable attention in evaluation studies, with papers: (1) presenting QCA as a complexity-appropriate method (Befani, 2013; Byrne, 2013; Verweij and Gerrits, 2013), (2) discussing methodological considerations and lessons learned (Befani et al., 2007; Pattyn et al., 2019; Sager and Andereggen, 2012; Thiem, 2014, 2017) and (3) using QCA to study the evaluation practices of different organizations or institutions (Holvoet and Dewachter, 2013; Pattyn, 2014; Van Voorst, 2017). Another CBM is cluster analysis. In evaluation studies, this approach is used to identify unique case types with different policy uptake profiles (Flygare et al., 2013; Gibson, 2003) identify internally homogeneous strata to sample for an intervention (Tipton, 2013) and identify distinct treatment subgroups who may exhibit heterogeneous intervention outcomes (Peck, 2005; Peck et al., 2012).

Scenario analysis is a planning method through which a small set of distinct but plausible scenarios are generated and analysed in regard to their implications for a current target of interest (Morell, 2005; Schwartz, 1991). Scenario analysis is a widely recognized in evaluation research (Buckley et al., 2015; Gates, 2016; Karani et al., 2015; Ling, 2003; Morell, 2005). By generating and analysing 'what if' scenarios, evaluators, policymakers and researchers can consider past counterfactuals or anticipate future challenges, reduce uncertainty by condensing the envelope of plausible outcomes, make assumptions about affected systems or

populations explicit, and conduct more thorough investigations of interventions into complex systems.

For example, Love and Russon (2000) developed multiple scenarios to probe the future of both international evaluation and evaluation societies and what these scenarios could mean for their development. More recently, several researchers have proposed incorporating scenario analysis into modelling approaches such as Bayesian Networks (Giffoni et al., 2018) or a cost benefit analysis framework (Campbell and Brown, 2005). These researchers parameterize input variables for their models to enable those using the models to explore alternative scenarios, reflecting different input parameters. For instance, Campbell and Brown (2005) discuss how a scenario that altered their original model by increasing initial costs for business investment would differentially impact stakeholder groups deciding on the investment. We build on this work in CBM and scenario simulation in the creation of CBSS.

Method: Case-based scenario simulation

CBSS is a social inquiry tool with potential for evaluation, grounded in a data-driven, case-based, computational systems modelling approach, which can be used to explore potential past counterfactuals and future projections of 'status quo' trends, and to test 'what if' scenarios related to some potential set of real or imagined interventions, given one's beliefs and assumptions about how those intervention(s) cause change. By *data-driven* we mean that the simulation environment is based on some set of real or simulated data. By *case-based*, we mean the focus is on the cases in a particular evaluation study and their complex profile of key factors (be they cross-sectional, pre-post, or longitudinal). By *computational*, we mean it uses the latest developments in machine intelligence and data visualization to create a simulated learning environment in which users can explore their theories, beliefs and assumptions of change from a complex systems perspective. The interventions or strategies explored can be, for example, a policy, a programme, or a set of individual or community-level goals.

The main advantages offered by CBSS are the ability to:

- Use advances in machine-driven cluster analysis to 'map' data to a topographical two-dimensional grid where cases are placed in proximity to similar cases – and from which, later scenarios are explored.
- Simplify the study population through k-means cluster analysis to identify the major and minor groups/trends around which the cases clustered.
- Identify differences in outcome by exploring cluster specific interventions, thereby allowing for the analysis of multiple possible solutions for a given population, including the analysis of unintended, unusable or impractical outcomes.
- Ground these different interventions (and their relative effectiveness) in changes to the complex set of factors (cluster profiles) upon which they are based, including their complex, nonlinear and (in terms of longitudinal data) dynamic interactions.
- Examine counterfactuals relative to a cluster-based solution
- Assess both short- and long-run effects of an intervention
- Run sensitivity tests on a set of interventions using Monte Carlo simulations, a method for testing outcomes under different ranges of variation.

Still, despite these advantages, CBSS is, at the end of the day, a model. And, as such, comes with its own built-in limitations, as do all computational models. Therefore, before

proceeding, a few caveats on causality in modelling are in order, all of which we have developed from Moore et al'.s (2019) recent article in *Evaluation*.

A complex systems approach to modelling: Key intellectual caveats

First, we need to be clear that no evaluation method, including CBSS, 'will ever be able to address the almost infinite number of uncertainties posed by the introduction of change into a complex system' (Moore et al., 2019: 36). However, adopting the type of case-based systems lens suggested by CBSS may help to 'drive the focus of evaluation (i.e. which of the multitude of uncertainties posed by interventions in complex systems do we need answers to in order to make decisions, or move the field forward)' (Moore et al., 2019: 36). It can also help to 'shape the interpretation of process and outcomes data' (Moore et al., 2019: 36).

Second, 'complex interventions in complex social systems', including exploring such strategies in CBSS 'pose almost infinite uncertainties and there will always be much going on outside of the field of vision of an individual study' (Moore et al., 2019: 37). 'However, a focus on discrete impacts of system change' as done with CBSS, 'does not necessarily betray a naïve view of how systems work, but may simply reflect a pragmatic focusing of research on core uncertainties' (Moore et al., 2019: 37). And this is, for us, one of the most powerful provisions that our approach provides, given its focus on interventions in complex clusters and their configurations, be such a study cross-sectional, pre-post, or longitudinal.

Third, we need to strongly emphasize that CBSS is a learning environment that aims to bridge the computational/quantitative/qualitative divide, as it requires users to be in direct and constant (i.e. iterative) interaction with the CBSS environment and their respective theories of change, be they sitting implicitly in the background of their minds or formally outlined and defined. For example, as Moore et al. (2019) state, 'Of course, it is never possible to identify all potential system level mechanisms and moderators of the effects of an evaluation', (Moore et al., 2019: 39), even in the case of CBSS:

> And no evaluation would be powered to formally model all of these. However, combining quantitative causal modelling [as in the case of CBSS] with qualitative process data [in the form of user-engagement with the learning platform of COMPLEX-IT] can play a vital role in building and testing theories about the processes of disrupting the functioning of complex social systems to optimize their impacts on health. (Moore et al., 2019: 39)

In other words, the goal here is not necessarily about identifying some underlying causal model, as much as it is about exploring and learning how various interventions might unfold for a given policy and the larger complex system in which it is situated. And such a goal, while humbler, is nonetheless very important.

How CBSS works

CBSS is implemented in a web-based platform named COMPLEX-IT which combines case-based modelling and scenario simulation into a streamlined tool for case-based analysis (Schimpf and Castellani, 2020). COMPLEX-IT is a 'thinking tool' that is intended to encourage iterative exploration of different case-based representations of data and cluster configurations, possible relationships between clusters, and the potential effects cluster profiles changes

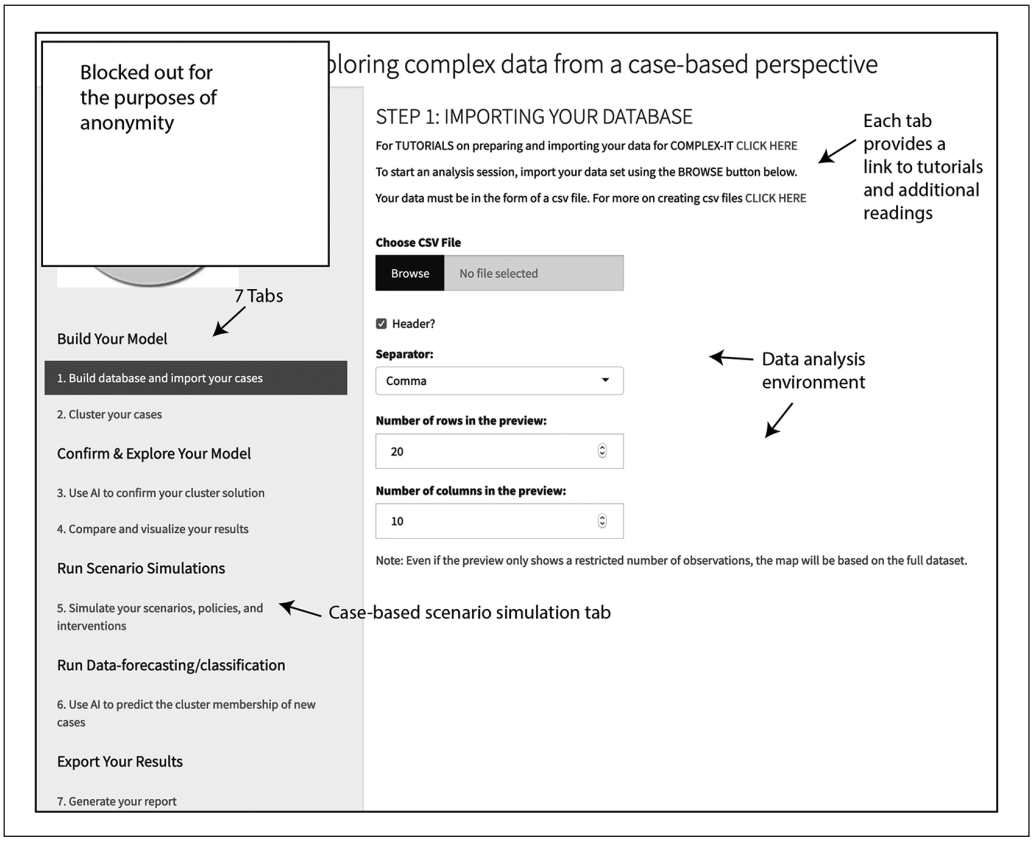

Figure 1. The COMPLEX-IT interface.

(i.e. alternate scenarios) might entail. COMPLEX-IT is shown in Figure 1. By combining methods into a single platform, employing visualizations to convey results and providing guiding questions COMPLEX-IT aims to support access to CBSS regardless of past experience. COMPLEX-IT can be accessed at (https://www.art-sciencefactory.com/complexit.html) which provides detailed tutorials and guidance for getting started. Generally speaking, CBSS works in two stages: data exploration and clustering cases through cluster analysis and machine learning; followed by scenario simulation, based on an exploration of the cluster solution.

Cased-based modelling. Before presenting k-means and the SOM as the central components of case-based modelling, a few general points need to be established. At its core, case-based modelling involves modelling complex systems or data as a collection of distinct case types with their own properties, trajectories and relationships with other cases. A case here is a holistic unit of study which is comprised of a 'profile' which contains elements representing internal attributes of the case, causal conditions, environmental factors, and other relevant elements. Following Byrne (2009), these cases are treated as complex systems. Cases, when viewed as complex systems, often exhibit regularities in their profiles such that several clusters may exist across the sample of cases (Castellani et al., 2013; Uprichard, 2009). Cluster analysis

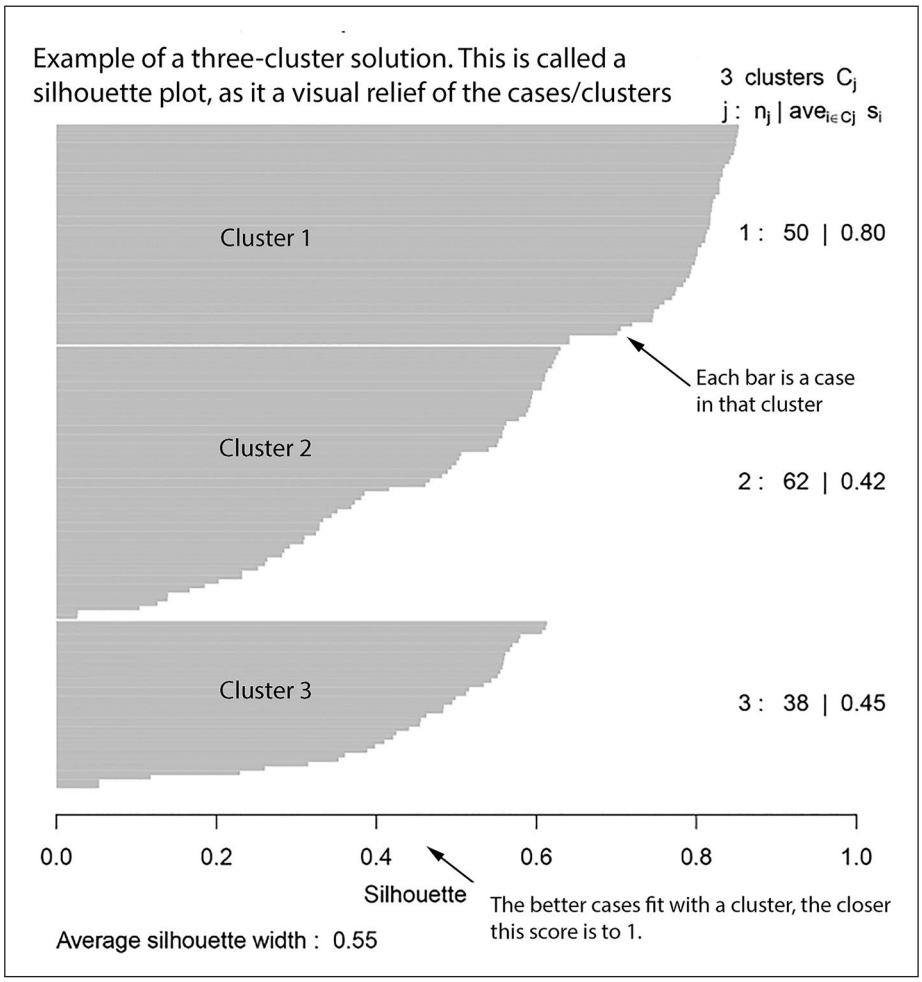

Figure 2. Silhouette plot of K-means clustering results.

techniques, including k-means and the SOM, seek to group cases into distinct subgroups/clusters. Thus, through case-based modelling an evaluator can reconceptualise a complex system/data experiencing an intervention(s) into a set of cases and seek to cluster those cases into major and minor profile groups (i.e. groups with many or few cases, respectively) to examine common trends, relationships within and across trends, and their environment and dynamics over time.

More specifically, as also shown in Figure 2, case-based modelling leverages k-means cluster analysis (Jain, 2010) to identify k clusters by placing cases within them, where k is defined by the evaluator, and where the algorithm seeks to minimizes the differences between a given case and the average values of its associated cluster. While k-means is most tightly integrated with CBSS, there is no reason other approaches to clustering cannot be used, including hierarchical clustering, partitional analysis, decision trees or QCA. The distinct advantage of QCA is that it is not variable based, focusing instead on

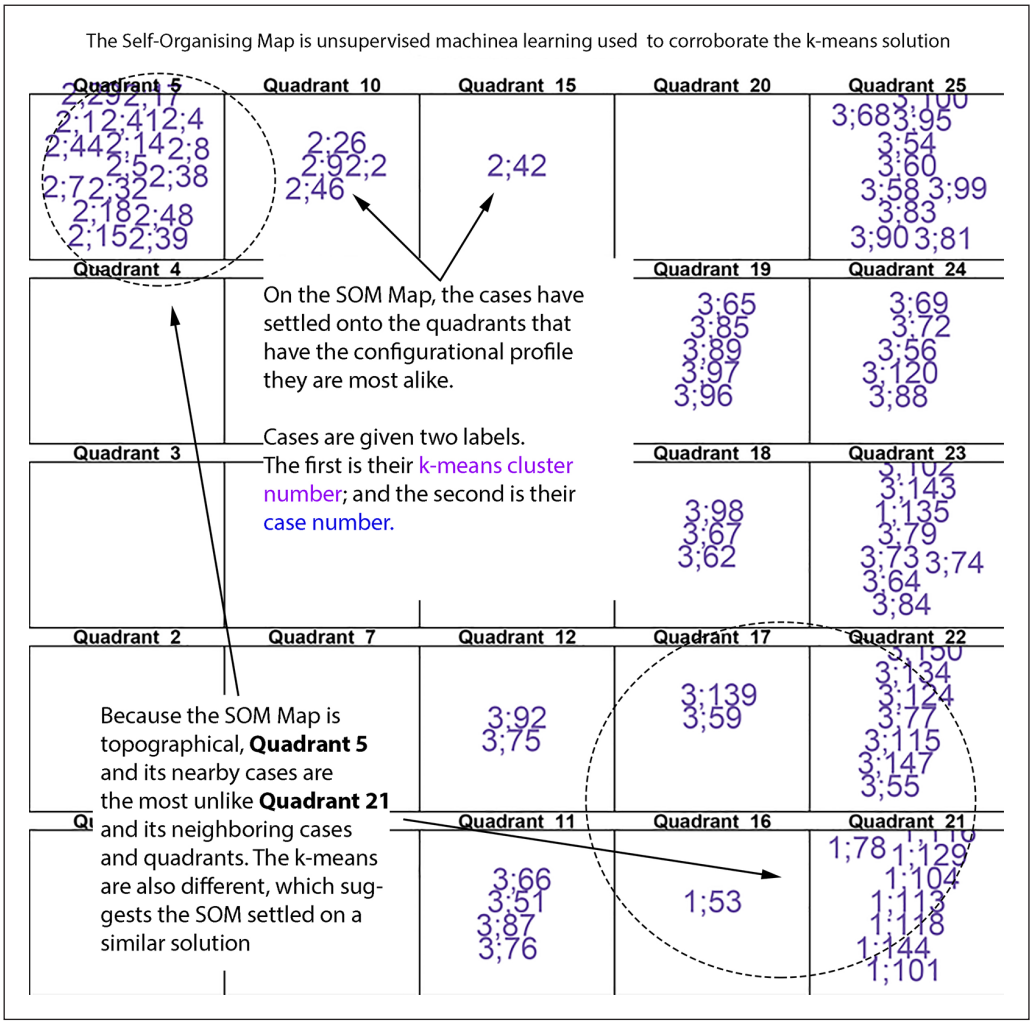

Figure 3. Self organized map and K-means corroboration.

different combinations of causal conditions that account for case outcome differences. While this difference in approach may sound minor, it is not, as QCA can help to arrive at otherwise difficult insights about the complex causal model(s) that accounts for cases and their differences (Befani et al., 2007).

Returning to case-based modelling, the case cluster/trends identified by k-means are then corroborated with the Kohonen 'self-organizing map' (SOM). The SOM operates by modelling a set of cases as a 2-dimensional abstract 'proximity map' where cases are closer to other cases with similar profiles and more distant from those with dissimilar profiles. More concretely, the SOM is an artificial neural network technique that uses an *n* x *m* grid where each grid has a profile matching the number of elements in a case profile (Kohonen, 2013). Cases are assigned to the grid node most like the case and that node and surrounding nodes 'learn' to be more like the assigned case. As shown in Figure 3, this leads to the mapping the cases across the grid such that

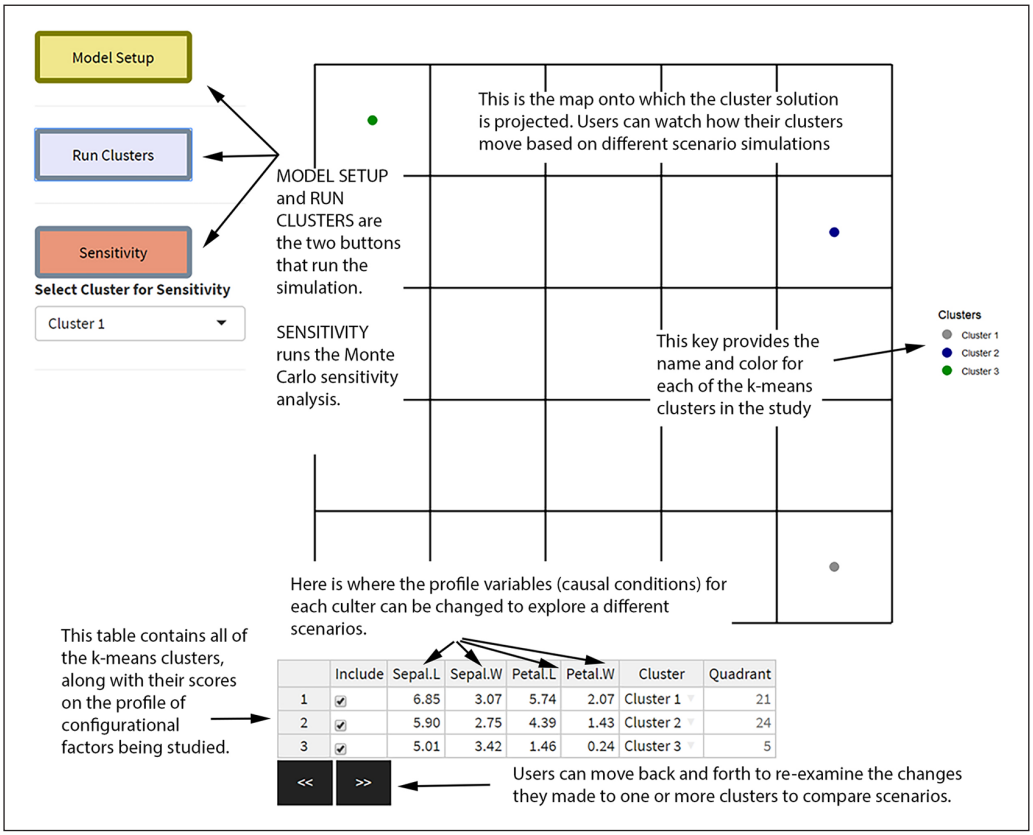

Figure 4. COMPLEX-IT scenario simulation.

similar cases are near each other and dissimilar cases are distant. After running this machine-driven SOM clustering, the evaluator can corroborate their k-means solution by comparing the k-means cluster profiles to grid node profiles and examining whether cases are assigned to similar profiles in both (Schimpf and Castellani, 2020). The SOM typically offers a finer-grained set of clusters which are determined by the neural network algorithm, in contrast to the evaluator setting *k* (i.e. number of clusters) in k-means. Comparing across cluster methods thus helps validate that the evaluator has landed on a reasonable representation of the underlying clusters in their data. In practice, multiple iterations of clustering at different *k* values and comparison with the SOM results is the intended use; this allows users to refine their understanding of the number of clusters that fit the data best, and which characteristics are key in determining clusters. It is worth noting that these approaches can be used with relatively small number of cases (*N*), for instance 15–20, as some evaluation studies have a small *N*.

Scenario simulation. With the clustering steps complete, the next step is scenario analysis/simulation. As we stated earlier, scenario analysis is a planning method used to generate possible future scenarios and to analyse their potential impact on a topic of interest (Börjeson et al., 2006; Morell, 2005; Schwartz, 1991). Bringing scenario analysis into case-based modelling follows advances in evaluation research in which alternative scenarios are explored through an empirically developed

Table 1. An overview of case studies used to demonstrate CBSS and COMPLEX-IT.

		Evaluation design	
		Cross-section	Longitudinal and/or quasi-experiment
Evaluation purpose	Learning and/or process evaluation	Heat Network Investments Project pilot evaluation (BEIS, 2018)	Frontline Fast-track Social work training pilot evaluation (control and intervention groups used) (DfE, 2016)
	Accountability and/or impact evaluation	Renewable Heat Incentive evaluation (BEIS, 2017)	The National Evaluation of Sure Start (DfE, 2008, 2010, 2012)

model by adjusting input parameters that represent different scenarios the modelled system or systems might encounter (Campbell and Brown, 2005; Giffoni et al., 2018).

Within CBSS, we call scenario analysis as 'scenario simulation'. As shown in Figure 4, the visualized model is the corroborated clusters representing the major/minor case trends while scenarios involve changes to cluster profiles and a simulation of how this change may affect the relative location of the affected cluster with respect to others. Said differently, an evaluator can make a hypothetical intervention on a k-means cluster solution they identified (which has been corroborated with and mapped to the SOM) and examine how this scenario results in the cluster gravitating towards or away from other SOM grid clusters. In essence, scenario simulations explore whether an intervention or change in a cluster's profile is sufficient to drive it towards or away from another cluster with desirable or undesirable qualities. (Note: as we outlined earlier in our caveats, scenario simulation does not make any assertion about the causality of changes in the cluster leading to its new position. Instead, it uses the SOM's re-positioning of a cluster as a means for exploring and prompting discussion on the movement of a cluster and potential causal relationships.)

Finally, as any change in a system is subject to uncertainty (Morell, 2005), hypothetical changes can be examined with a Monte Carlo simulation (See Figure 4), which allows user changes to a cluster to take a band of values depending on their judgement of the precision of the change, that is, for sensitivity analysis. Results from the Monte Carlo display the distribution of cluster profiles the changed cluster lands on based on multiple analyses with different values across the uncertainty band. Thus, collectively case-based modelling and scenario simulation form CBSS, provides a novel way to model and explore complex systems and hypothetical changes in said systems.

Case studies

In this section, we review four published evaluation studies to demonstrate how CBSS and its platform COMPLEX-IT can support evaluation in practice. To demonstrate the utility of our approach to a wider audience (as opposed to a single in-depth example) our case studies (as shown in Table 1) come from a range of domain settings, evaluation contexts, and approaches, and are organized according to (1) evaluations with either a learning or process evaluation purpose or those with an accountability or impact evaluation purpose and (2) evaluations with a cross-sectional design, or those with a longitudinal and/or quasi-experiment design.

This approach, however, did come with some costs. While our driving goal for selecting studies was to provide examples that were applicable to a broad audience of

evaluators and evaluation researchers, most evaluations do not publish or make their data publicly available. This is because there are often severe restrictions on what may be published for several valid reasons. These reasons include, but are not limited to, participant consent being collected for specific evaluation goals which may not translate to more general research goals, unavailability of raw data due to anonymity concerns and some data may require renegotiating consent. We therefore chose to describe the potential use of CBSS for our four studies in acknowledgement of limitations of acquiring detailed data and in order to cover a more comprehensive set of evaluation design and purpose circumstances. This likewise enabled us to show the versatility of the proposed approach across evaluation circumstances.

Heat Network Investments Project pilot evaluation

Our first published case study is an evaluation of the UK government's 'Heat Network Investment Project' (HNIP) pilot. The HNIP provides capital support (e.g. loans and grants) for applicants for the development of heat networks (i.e. the transfer of hot water for space and water heating, across different buildings) in England and Wales. It is planned to be run from 2016 to 2021. The pilot was designed to generate learning to be used in the main scheme, and was only open to public-sector applicants. It ran from 2016 to 2017, and its evaluation was conducted from 2016 to 2018. The evaluation was published by the UK Department for Business, Energy and Industrial Strategy (BEIS, 2018).

The evaluation focussed on the following: the successes and failures of the pilot; any patterns in the successful or unsuccessful applicants; views of stakeholders on the heat networks market and their views on barriers and preferred financial options; and stakeholder views of the HNIP pilot. The evaluation was theory-based and used interviews, documentary analysis, and basic numerical analysis of the application data. Note that the evaluation of the full programme is presently (circa 2020) ongoing.

Cluster analysis and embracing complexity. As depicted in Figures 2 and 3 earlier, one of the ways CBSS/COMPLEX-IT could complement this study is by using cluster analysis (both k-means and the SOM) to identify the complex profile of factors that distinguished successful and unsuccessful applicants – without worry to the low n of the report. As the numerical analysis in BEIS (2018) shows, only one factor was considered at a time in relation to successful and unsuccessful applications (e.g. technology type, customers, size and scale of funding sought). In contrast, because CBSS embraces complexity and does not reduce explanation, it would have allowed the evaluators to consider multiple factors (and their complex intersection) simultaneously, to arrive at a more causally complex understanding of why certain applications were more successful than others.

The sensitivity of the data on applications also means that all reporting had to be carefully checked so that no applicants could be identified from the results; COMPLEX-IT would be helpful in this regard, as reporting could use the cluster averages without needing to provide details of who or how many individual applicants.

Machine learning and data forecasting. Also, given that COMPLEX-IT uses the SOM, which incorporates machine learning, the results in regard to successful and successful applicants

could be treated as a training dataset from which the outcome of future applicants could be predicted or forecasted.

Data mining/exploration. Taking a more speculative view the COMPLEX-IT could also be used to open up the potential for different evaluation questions and foci. For example, as the policy moves from the pilot to the main scheme, the evaluators and client may find it useful to explore how using a different scoring mechanism for applicants could affect outcomes and funding decisions. The funding decisions were made by scoring various elements of applicants' proposals; these could be scored with different rubrics, different thresholds for values, and/or with whole different sections or partitioning of elements.

Scenario simulation. COMPLEX-IT, through simulating scenarios provides another way of visually exploring new evaluation questions in regard to applicant outcomes. This could be done by using the original scoring data to generate the clusters, then either re-running the analysis with different scoring data (e.g. say with an additional factor giving a score for some new element of the proposals), or, using a scenario, and changing the scoring values at the cluster level (as shown in the bottom of Figure 4), that is, simulating assessing one or several clusters more harshly on a specific element. The results CBSS/COMPLEX-IT would push back at us (i.e. different clusters, more or less clusters) and allow us to consider what effect changing the assessment process may have.

Renewable Heat Incentive evaluation

The Renewable Heat Incentive (RHI) is a payment system for the generation of heat from renewable sources. Begun in November 2011, it replaced the Low Carbon Building Programme (for non-domestic applications), and was then extended to domestic application too in April 2014 (in effect replacing the Renewable Heat Premium Payment). The RHI is designed to support households, businesses, public bodies and charities in transitioning from conventional forms of heating to renewable alternatives. There are multiple evaluation reports on the RHI published by BEIS, split by the domestic and non-domestic scheme and through time.

For our study, we will focus on the synthesis report published as BEIS (2017). BEIS (2017) highlights how the evaluation of the RHI has focussed on three main types of actors: applicants (broken down into smaller subgroups, including domestic and non-domestic), installers of renewable heat technology, and non-applicants. Various types of data have been collected on each of these using a range of research methods and available administrative data. As with the HNIP, it is clear COMPLEX-IT could be used to explore any patterns and potential clusters in any of these groups. Beyond this task, one of the key questions the evaluation of the RHI has explored was whether there are many non-applicants that could or should have been applicants (from the point of view of the policy's aims); and if so, why they did not apply.

Cluster analysis. Clustering could be run on applicants and non-applicants together to search for any clusters of non-applicants that are similar to applicant clusters. This process could identify which factors or combinations of factors differentiate these 'nearly-applicants' from actual applicants.

Scenario simulation and counterfactuals. Scenario simulation could be used to explore how changing key combinations of variables (causal conditions) in their profile may have led them to become successful applicants. What is critical, in terms of counterfactuals is that changing key combinations of variables may reveal how more than one set of changes could lead to the same results. Moreover, the degree of change needed on one or more factors in order to move a non-applicant towards applicant clusters could also reveal the robustness of non-applicant clusters.

Thinking about complex causality. In short, this approach to scenario simulation would have the evaluators focused on all of the factors in the study and their complex interaction. This is key, CBSS is a way to embrace and preserve the complexity of the topic being studied, rather than reducing the evaluation to a small number of factors. The results could then be vetted against previous evaluation studies or data. For example, the simulation results could be critically assessed by the extensive research already conducted in the evaluation on customer journeys (i.e. exploring and understanding how customers go from not having any renewable heat source, to getting one, and applying to the RHI or not). Findings on customer journeys could be used to make sense of the profile elements which the clustering suggests distinguishes applicants and non-applicants, and/or identify which profile elements are plausible to change at the cluster level in the scenario simulation.

Frontline Fast-Track Social Work training programme pilot evaluation

The Frontline Fast-Track Social work training programme aims to recruit future social workers who exhibit both high academic potential and strong interpersonal skills. The educational model for recruits has a heavy emphasis on practice-based learning and was geared specifically to help students work with at-risk children. The programme was piloted in Greater Manchester and Greater London from 2013 to 2015 and was expanded to the North-East until the programme ended in 2017. The pilot was used to understand more about the recruits the programme attracted and the learning outcomes it evinced.

An evaluation of the programme was published in DfE (2016). It focused its evaluation questions on (1) whether the programme attracted high-quality candidates, (2) the quality of the educational programme delivered, and (3) and a quasi-experimental assessment of the educational gains of Frontline participants in comparison to other educational programmes. The evaluation used aggregate education data, surveys of Frontline and comparison students, interviews and focus groups with various stakeholders, and a performance assessment of Frontline and comparison students social work skills. While the quasi-experimental assessment conducted with Frontline trainees and students from other social work programmes revealed that Frontline trainees outperformed other students on simulated social work interviews with service users, this analysis relied on simple performance averages for comparison.

Cluster analysis. A more nuanced analysis could use COMPLEX-IT to identify clusters with different performance profiles, drawing on several metrics by which students' performance was assessed, from both video of their interaction with the 'client' and a post-interview written reflection, for Frontline participants and comparison students. By comparing the performance profiles of these clusters, it may be possible to identify some distinguishing or configurational

patterns the experimental and comparison group. Comparisons could also uncover: (1) any clusters that performed similarly across the groups, (2) Frontline clusters that performed worse than comparison clusters, and (3) the unique performance of Frontline students to the other groupings. This could aid evaluators in understanding the effect of the Frontline pilot and what types of learning gains students may have achieved vis-a-vis traditional social work programmes. It could also provide insight into the how Frontline's emphasis on practice shapes learning.

Understanding alternative outcomes. More hypothetically, the evaluators may want to explore the clusters of Frontline and comparison students exclusively on their writing scores, in light of evaluators additional analysis revealing that the groups showed less difference on this outcome. This would involve reforming the clusters to only include assessment items from student reflections and then comparing across groups. As mentioned above, the Frontline programme placed a heavy emphasis on learning through social work practice, so closer examination of the differences between student groups on more traditional academic outcomes like writing may provide greater insight into how the configuration of the programme led to the learning outcomes observed and inform future programmes based on this pilot.

Scenario simulation. In a similar hypothetical vein, evaluators may wish to explore alternative scenarios, such as positively adjusting non-Frontline student performance because in the original study evaluators discovered that Frontline students had better academic track-records than comparison group students. This may help better understand the effect of the programme by creating a more level comparison.

The national evaluation of sure start

Sure Start is a UK government initiative intended to enhance the life chances of young children growing up in disadvantaged neighbourhoods. It began in earnest in 1999 and is ongoing, though it has gone through significant changes since its start. The initiative used an area-based approach with all young children and their families in a certain area being the 'targets' of the intervention. Each area had a high level of autonomy on what sets of services they wished to provide or improve as part of the initiative. However, they were intended to cover certain core services such as: outreach and home visiting; support to families and parents; support for good quality play, learning and childcare; and primary and community healthcare and support for children and parents with special needs.

The National Evaluation of Sure Start (NESS) (DfE, 2008, 2010, 2012) ran from 2003 to 2011 and was focussed on evaluating if children and families benefitted from Sure Start (and if so, how and for whom, under what conditions). It used a longitudinal and quasi-experimental design comparing children/ families in similar areas receiving and not receiving Sure Start programmes. For families not in programmes, a subset of the Millennium Cohort Study was used.

In regards to the study, NESS took a traditional quantitative impact evaluation focus, assessing impact using 15 different outcome measures (in the final 2012 phase) and 8 which were measured throughout the evaluation. Methodological issues were discussed at a reasonable length, though these mostly concern they ways in which the study overcame a range of

issues that forced them away from being able to conduct a traditional Randomized Control Trial (i.e. sampling methods, statistical analyses used).

The NESS was a controversial evaluation for a number of reasons, not least caused by the prominence of the policy. Lloyd and Harrington (2012), reflecting on the NESS, outline a range of evaluation challenges it faced, and conclude that one of the most important was the disconnect between local and national levels of the evaluation, with over-reliance on the national level, and inadequate connections to local evaluation teams and evidence. We do not wish to wade into these debates here, as it is beyond the scope and focus of this paper; however we use the published NESS report to demonstrate our argument as it is still applicable for exploration, but note that these reports were critiqued at the time.

Thinking in complex realist terms. Epistemologically speaking, we could assume that a quasi-experimental approach to evaluation is directly at odds with a case-based approach. However, it is our assertion that the use of these approaches in combination should always be considered if at all possible and appropriate (CECAN, 2018; Stern, 2015). For example, NESS makes efforts to consider the impacts of Sure Start on sub-populations and notes,

> results for sub-populations can be as important as those for the total population . . . [t]his is increasingly important as children's centre services are increasingly targeted at the most vulnerable, and also service delivery may be targeted differently for specific sub-populations. (DfE, 2012)

In effect, this consideration of subpopulations is a first step towards a case-based approach. CBSS would allow a study like NESS to take this further.

Cluster analysis and scenario simulation. For example, clustering would allow evaluators to look for subgroups based on patterns in the data, rather than using prior beliefs. In order to separate subgroup variations in the control and experiment samples and potential heterogeneity in intervention impact, CBSS or other clustering approaches can be used before traditional statistical analysis (see Peck, 2005). Using it before can complement other statistical analysis by helping refine an evaluator's understanding of their data, inform some of the theory they are using to structure their analysis or identify subgroups and explore intervention effects on said subgroups. Or if an evaluator is more interested in outcomes it could be used after the analysis. Once we have examined for any statistically significant differences between the control and intervention groups, and perhaps subgroups within them, we can then explore the data for post-intervention outcome clusters using the clustering methods, or explore the robustness of any associated outcome clusters using the scenario simulation.

Longitudinal cluster analysis. While we have not yet discussed it much, COMPLEX-IT is also well-placed to be used on longitudinal data such as that collected in the NESS. Here, there are two basic options; to use longitudinal clustering based on absolute values (i.e. factor values for various time points) and their change through time, or to cluster on relative rates of change (i.e. time period 2 minus time period 1 values) through time. In the NESS, using COMPLEX-IT to do this would likely have been most interesting to explore groupings based on relative change, as the existing analysis already covers absolute change.

Discussion

As we hopefully demonstrated, across the four case studies a number of themes emerged, some we only lightly mentioned, while others require more elaboration. We outline these additional points below. First, in terms of technique, the clustering and scenarios techniques in CBSS showed wide-ranging potential for enabling greater exploration of the data evaluators have, helping them understand the evaluation context, contributing new understanding, and raising new questions or evaluation directions across these studies.

Second, the cases demonstrate that CBSS is not at odds with other methods and evaluation data. Instead, it works well with them. For example, the customer journeys collected in the RHI study could be used to help interpret differences in the cluster profiles of applications and non-applications or inform which profile elements to change in scenarios.

Third, the synergy between CBSS and other methods used in the evaluation studies reviewed above highlights a broader point about the importance of not restricting evaluation to a single method or set of methods. As Stern (2015: 9) writes, 'No one methodological approach is best or even sufficient on its own, which is why we need to draw on a broad range of approaches and methods for impact evaluation'. When evaluating interventions into complex systems multiple methods are often required to render visible changes in a dynamic context (CECAN, 2018); in short, a pluralist approach is often needed (Stern, 2015). Reviewing these studies shows that CBSS has the versatility to integrate with other methods and extend evaluation insights.

Fourth, it is also noteworthy that while methods like clustering have potential for all the studies reviewed, what CBSS brought to each study differed. For example, CBSS supported: (1) richer descriptions and holistic characterizations of populations targeted by the programmes reviewed (e.g. HNIP); (2) in-depth comparisons of similarities and differences across distinct groups (e.g. Frontline and Sure Start); (3) the opportunity to explore counterfactuals or 'what if' scenarios (e.g. HNIP and Frontline); and (4) examining the strength of differences between clusters representing different case types (e.g. RHI and Sure Start).

Fifth, and relative to the last point, given the dynamic nature, context sensitivity, and path dependency of complex systems, CBSS provides a broader understanding of the possibility-space around a given system of study (Byrne, 2013; Moore et al., 2019). For example, it was possible to craft different scenarios across the four published studies that either generated insight into the policy context or helped determine the degree of impact an intervention had. In other words, CBSS allows evaluators to examine different scenarios in a low-risk, low-cost simulated environment, to best address their evaluation goals. As evidence, both the Sure Start and RHI evaluation, which have accountability goals, presented scenarios that evaluated the ease at which different groups could be nudged to become associated with other desirable or undesirable clusters. Furthermore, these scenario results could be viewed as a measure of outcome stability and thus more directly speak to accountability goals.

Sixth, depending on the nature of the evaluation study, CBSS also allows for new evaluation directions or questions to be addressed. For instance, in the Frontline study there were two distinct directions. The first sought to explore differences between experiment and comparison groups on an outcome that was underemphasized by the Frontline programme while the second sought to make a more level comparison between groups. Evaluators may be interested in both, one, neither, or other complementary questions; what is central is that CBSS can open multiple such inquiries for further exploration.

Seventh, a more general point touched on throughout is that by supporting the discovery and analysis of unique clusters, their dynamics, and their relationship with other clusters, CBSS is well aligned with a realist programme of evaluation. In particular, CBSS enables evaluators to identify distinct subgroups within their data, their unique properties and examine the differential effect interventions may have on these groups to uncover 'what works for whom in which circumstances' (Pawson and Tilley, 1997: 77).

Eighth, because of its case-based approach, applying CBSS to evaluations raises two critical methodological considerations. First, what are the cases should be evaluated? Charles Ragin, the creator of QCA, discusses this issue of 'casing' extensively (see Ragin, 2009; Ragin and Becker, 1992). Ragin notes that 'casing' in QCA (and applicable to other CBMs), uses a realist perspective in that cases are considered real entities, but identifying the most relevant cases including encompassing unit and profile is an inherently iterative process which unfolds over the analysis cycle (Ragin, 2009: 524). Likewise, using CBSS may invite questions about whether the evaluator has chosen appropriate cases, profile elements and/or units of analysis. The nature of cases in evaluation is not a simple question; instead it emerges from the analysis and through a dialogue with the data and context of the study. This process of case identification and selection is further shaped by the practicalities of case availability or accessibility. COMPLEX-IT supports evaluators to quickly conceptualize, configure and test cases (both through human and machine-driven cluster comparison and against other evaluation data or theory). However, if a decision to change the unit of analysis is made, this may require additional data collection which will be more costly. Of note, this conceptualization, configuration and testing cycle is also one of the key ways in which CBSS and its platform COMPLEX-IT embody a learning environment through which evaluators can develop their working models and casual explanations and/or theories of change.

The second consideration pertains to an assumption made in CBM and consequently CBSS: cases and their profile specifications are reasonable approximations of their real-world counterparts. This assumption may not hold for several reasons and warrants different responses depending on the misalignment. One common possibility is that the case profile elements are poorly specified, contain spurious or redundant elements or have other limitations. Alternatively, the units being represented as cases may be too large or too small and need to be adjusted accordingly. When cases are poorly specified or of an inappropriate scope for the study, results will be greatly hampered and unreliable. Therefore, following the SACS toolkit guidance on case-based modelling, evaluators should conduct regular validity checks (Castellani and Hafferty, 2009: 79–80) to step back and re-examine their case modelling, whether it is internally consistent and consistent with other study aspects and/or relevant theories. These validity checks give rise to iterating on the study's cases, questions, and analysis as discussed above.

In a similar vein, another problem may arise where either the evaluator cannot find an adequate cluster representation or a relatively stable cluster representation emerges, however they disagree with other evaluation results, relevant theories or experience with the system(s) under study. This may also emerge when exploring the scenario simulation leads to unusual or contradictory outcomes. Under these circumstances, it may be that elements in the case profiles poorly capture critical causal factors that influence case dynamics and differences between cases behaviours. One way to address this challenge is to draw on theories of change (Coryn et al., 2011; Ofek, 2017; Rolfe, 2019; Wilkinson et al., 2021), systems maps (Barbrook-Johnson, 2020; Barbrook-Johnson, 2019; Barbrook-Johnson and Penn, 2021) or other theories

that demarcate the major causal forces in and around the cases. These theories can then be used to inform how cases are built and what factors are manipulated in scenarios. COMPLEX-IT also offers links to system mapping software to support this (see: https://www.art-sciencefactory.com/complexit.html). In general, it is a good practice to draw on relevant theories or experience with the system(s) being modelled to guide case construction and identify plausible scenarios to explore.

Another related possibility is that the data may have a lot of mismeasurement, inaccuracy, unevenness in quality or completeness, weak relevance, or simply poorly capture aspects of the system(s) for modelling. If data collection is finished and there are no alternative data sources, other methods that can take advantage of the data may be preferable. However, if the primary issue revolves around unevenness in completeness of cases, it may be possible conduct some analysis if a sufficient number of cases are more complete or of a higher quality. Regular validity checks throughout CBSS will also help in identifying these challenges. Aside from the aforementioned possibilities, it may turn out that cases themselves are not fundamentally complex. Cases may only have a few meaningful attributes or most or all of their attributes may have simple or linear distributions. Under these circumstances it may be unnecessary, although not inherently wrong, to draw on CBSS or other CBM as linear statistics and associated techniques are well suited for analysing such situations.

Conclusion

To aid the growing research programme of complexity appropriate methods for evaluation studies, we introduce CBSS to advance the use of CBMs in evaluation. Furthermore, we introduced COMPLEX-IT, that increases CBSS accessibility and facilitates evaluators ability to iteratively explore their data (Schimpf and Castellani, 2020).

We next reviewed evaluation research on CBMs and scenario analysis to show how this work builds on past research. We then presented a prospective view of what CBSS could bring to four previously published evaluation studies. Note the studies were not meant to be exhaustive of what CBSS and COMPLEX-IT could bring to evaluation studies but sought to illustrate their use across different evaluation purposes and designs.

CBSS methods showed synergy with other evaluation methods and findings. This synergy also highlighted an important insight: rarely does a single method provide all the answers needed in an evaluation; therefore, a pluralist approach is often needed (Stern, 2015), which CBSS and COMPLEX-IT are well positioned to support.

Reviewing across the evaluation studies also indicated that CBSS and COMPLEX-IT could contribute different questions, insights and directions depending on the nature of the study. Finally, several methodological points about CBMs were discussed including the issue of deciding on appropriate cases, (i.e. casing, see Ragin, 2009; Ragin and Becker, 1992), conducting validity checks on an evaluators model (Castellani and Haffery, 2009: 79–80) and incorporating theories of change (Coryn et al., 2011; Ofek, 2017; Rolfe, 2019), system maps (Barbrook-Johnson, 2020; Barbrook-Johnson, 2019; Barbrook-Johnson and Penn, 2021) or other theories to inform case construction and scenario exploration.

In short, we believe CBSS and COMPLEX-IT hold potential value for evaluators by adding to the growing landscape of complexity-appropriate methods in evaluation and offering an approach that is highly synergistic with other methods and which can flexibly complement different evaluation purposes and open several new avenues of exploration as demonstrated in

the reviewed evaluation studies. CBSS likewise aligns well with the realist evaluation agenda by enabling evaluators to discover and cluster unique case types and explore their dynamics, relationships, environmental influences and responses to interventions. COMPLEX-IT, in particular, offers a tool that enables evaluators to quickly engage in case-based modelling and scenario simulation regardless of prior experience. This is especially timely in light of the ongoing challenges evaluators face including limited-time to try new methods, work-contexts that are often less supportive of new methods, and evaluators perceptions of required data, that may then constrain their use of new methods (Barbrook-Johnson et al., 2019); COMPLEX-IT's increased accessibility can help alleviate the first two barriers, and its support for exploration and iterative analysis can help with the third barrier.

Considerable work remains on developing methodological innovations for modelling and evaluating interventions in complex systems. CBSS and its platform COMPLEX-IT make strides in this direction by enabling evaluators to probe complex systems, generate new ideas, questions and directions in evaluation, draw distinctions between heterogeneous intervention effects and complement a wide-array of methods and evaluation designs.

Acknowledgements

The authors with to thank the Centre for Evaluation Across the Nexus (CECAN) for their intellectual and financial support and Durham University for its financial support.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Economic and Social Research Council [grant numbers ES/N012550/1 and ES/S000402/1] and from Durham University via the ESRC pathways to impact grant.

ORCID iDs

Corey Schimpf https://orcid.org/0000-0003-2706-3282

Pete Barbrook-Johnson https://orcid.org/0000-0002-7757-9132

References

Barbrook-Johnson P (2019) Negotiating complexity in evaluation planning: A Participatory Systems Map of the Energy Trilemma. CECAN EPNN No. 12. Available at: www.cecan.ac.uk/resources

Barbrook-Johnson P (2020) A Participatory Systems Mapping in action: Supporting the evaluation of the Renewable Heat Incentive. CECAN EPNN No. 17. Available at: www.cecan.ac.uk/resources

Barbrook-Johnson P, Schimpf C and Castellani B (2019) Reflections on the use of complexity-appropriate computational modeling for public policy evaluation in the UK. *Journal on Policy and Complex Systems* 5(1): 55–70.

Barbrook-Johnson P and Penn AS (2021) Participatory systems mapping for complex energy policy evaluation. *Evaluation* 27(1): 57–79.

Befani B (2013) Between complexity and generalization: Addressing evaluation challenges with QCA. *Evaluation* 19(3): 269–83.

Befani B, Ledermann S and Sager F (2007) Realistic evaluation and QCA: Conceptual parallels and an empirical application. *Evaluation* 13(2): 171–92.

Börjeson L, Höjer M, Dreborg K-H, et al. (2006) Scenario types and techniques: Towards a user's guide. *Futures* 38(7): 723–39.

Bowering DJ (1984) Impact analysis using an integrated data base: A case study. *New Directions for Evaluation* 1984: 73–110.

Buckley J, Archibald T, Hargraves M, et al. (2015) Defining and teaching evaluative thinking: Insights from research on critical thinking. *American Journal of Evaluation* 36(3): 375–88.

Byrne D (2009) Complex realist and configurational approaches to cases: A radical synthesis. In: Byrne D and Ragin CC (eds) *The SAGE Handbook of Case-Based Methods*. Thousand Oaks, CA: SAGE, 101–11.

Byrne D (2013) Evaluating complex social interventions in a complex world. *Evaluation* 19(3): 217–28.

Byrne D and Ragin CC (eds) (2009) *The SAGE Handbook of Case-Based Methods*. Thousand Oaks, CA: SAGE.

Campbell HF and Brown RPC (2005) A multiple account framework for cost–benefit analysis. *Evaluation and Program Planning* 28(1): 23–32.

Castellani B, Barbrook-Johnson P and Schimpf C (2019) Case-based methods and agent-based modelling: Bridging the divide to leverage their combined strengths. *International Journal of Social Research Methodology* 22(4): 403–416.

Castellani B and Haffery F (2009) *Sociology and Complexity Science: A New Field of Inquiry*. Germany: Springer.

Castellani B, Schimpf C and Hafferty F (2013) Medical sociology and case-based complexity science: A user's guide. In: Sturmberg JP and Martin CM (eds) *Handbook of Systems and Complexity in Health*. New York: Springer, 521–535.

CECAN (2018) Policy evaluation for a complex world. *CECAN Report*. Available at: www.cecan.ac.uk/resources

Coryn CLS, Noakes LA, Westine CD, et al. (2011) A systematic review of theory-driven evaluation practice from 1990-2009. *American Journal of Evaluation* 32(2): 199–226.

Cousins JB and Bourgeois I (2014) Cross-case analysis and implications for research, theory, and practice: Cross-case analysis and implications for research, theory, and practice. *New Directions for Evaluation* 2014(141): 101–19.

Department for Business, Energy and Industrial Strategy (BEIS) (2017) RHI synthesis report. Available at: https://www.gov.uk/government/collections/renewable-heat-incentive-evaluation

Department for Business, Energy and Industrial Strategy (BEIS) (2018) Evaluation of the Heat Networks Investment Project (HNIP): Pilot process report. Available at: https://www.gov.uk/government/publications/evaluation-of-the-heat-networks-investment-project-hnip-pilot-scheme

Department for Education (DfE) (2008) The impact of Sure Start Local Programmes on three year olds and their families. Available at: https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=6473

Department for Education (DfE) (2010) The impact of Sure Start Local Programmes on five year olds and their families. Available at: https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=6473

Department for Education (DfE) (2012) The impact of Sure Start Local Programmes on seven year olds and their families. Available at https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=6473

Department for Education (DfE) (2016) Independent evaluation of the Frontline pilot. Available at: https://www.gov.uk/government/publications/frontline-pilot-independent-evaluation

Flygare E, Gill PE and Johansson B (2013) Lessons from a concurrent evaluation of eight antibullying programs used in Sweden. *American Journal of Evaluation* 34(2): 170–89.

Garaway G (1996) The case-study model: An organizational strategy for cross-cultural evaluation. *Evaluation* 2(2): 201–11.

Gates EF (2016) Making sense of the emerging conversation in evaluation about systems thinking and complexity science. *Evaluation and Program Planning* 59: 62–73.

Gibson CM (2003) Privileging the participant: The importance of sub-group analysis in social welfare evaluations. *American Journal of Evaluation* 24: 443–69.

Giffoni F, Salini S and Sirtori E (2018) Evaluating business support measures: The Bayesian Network approach. *Evaluation* 24(2): 133–52.

Gilbert N, Ahrweiler P, Barbrook-Johnson P, et al. (2018) Computational modelling of public policy: Reflections on practice. *Journal of Artificial Societies and Social Simulation* 21(1): 1–14.

Hayton K (2015) Evaluating the evidence – A move to more realistic evaluation: A case study of Regional Selective Assistance in Scotland. *Evaluation* 21(2): 248–62.

Holvoet N and Dewachter S (2013) Multiple paths to effective national evaluation societies: Evidence from 37 low- and middle-income countries. *American Journal of Evaluation* 34(4): 519–44.

Jain AK (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31(8): 651–66.

Koenig G (2009) Realistic evaluation and case studies: Stretching the potential. *Evaluation* 15(1): 9–30.

Karani I, Mayhew J and Anderson S (2015) Tracking adaptation and measuring development in Isiolo County, Kenya. *New Directions for Evaluation* 2015(147): 75–87.

Kohonen T (2013) Essentials of the self-organizing map. *Neural Networks* 37: 52–65.

Leone L (2008) Realistic evaluation of an illicit drug deterrence programme: Analysis of a case study. *Evaluation* 14(1): 9–28.

Ling T (2003) Ex ante evaluation and the changing public audit function. *Evaluation* 9(4): 437–452.

Lloyd N and Harrington L (2012) The challenges to effective outcome evaluation of a national, multi-agency initiative: The experience of sure start. *Evaluation* 18(1): 93–109.

Love AJ and Russon C (2000) Building a worldwide evaluation community: Past, present, and future. *Evaluation and Program Planning* 23(4): 449–59.

Mookherji S and LaFond A (2013) Strategies to maximize generalization from multiple case studies: Lessons from the Africa Routine Immunization System Essentials (ARISE) project. *Evaluation* 19(3): 284–303.

Moore GF, Evans RE, Hawkins J, et al. (2019) From complex social interventions to interventions in complex social systems: Future directions and unresolved questions for intervention development and evaluation. *Evaluation* 25(1): 23–45.

Morell JA (2005) Why are there unintended consequences of program action, and what are the implications for doing evaluation? *American Journal of Evaluation* 26(4): 444–63.

Ofek Y (2017) An examination of evaluation users' preferences for program and actor-oriented theoreis of change. *Evaluation* 23(2): 172–91.

Pattyn V (2014) Why organizations (do not) evaluate? Explaining evaluation activity through the lens of configurational comparative methods. *Evaluation* 20(3): 348–67.

Pattyn V, Molenveld A and Befani B (2019) Qualitative comparative analysis as an evaluation tool: Lessons from an application in development cooperation. *American Journal of Evaluation* 40(1): 55–74.

Pawson R and Tilley N (1997) *Realistic Evaluation*. London: SAGE.

Peck LR (2005) Using cluster analysis in program evaluation. *Evaluation Review* 29(2): 178–96.

Peck LR, D'Attoma I, Camillo F, et al. (2012) A new strategy for reducing selection bias in nonexperimental evaluations, and the case of how public assistance receipt affects charitable giving: Reducing selection bias in nonexperimental evaluations. *Policy Studies Journal* 40(4): 601–25.

Ragin CC (2009) Reflections on casing and case-oriented research. In: Byrne D and Ragin CC (eds) *The SAGE Handbook of Case-Based Methods*. Thousand Oaks, CA: SAGE, 1–13.

Ragin CC and Becker HS (eds) (1992) *What Is a Case? Exploring the Foundations of Social Inquiry*. Cambridge: Cambridge University Press.

Rolfe S (2019) Combining theories of Change and Realist Evaluation in practice: Lessons from a research on evaluation study. *Evaluation* 25(3): 294–316.

Sager F and Andereggen C (2012) Dealing with complex causality in realist synthesis: The promise of qualitative comparative analysis. *American Journal of Evaluation* 33(1): 60–78.

Savaya R, Spiro S and Elran-Barak R (2008) Sustainability of social programs: A comparative case study analysis. *American Journal of Evaluation* 29(4): 478–93.

Schimpf C and Castellani B (2020) COMPLEX-IT: A case-based modelling and scenario simulation platform for social inquiry. *Journal of Open Research Software* 8(1).
Schwartz P (1991) *The Art of the Long View: Planning for the Future in an Uncertain World*. New York: Doubleday.
Stern E (2015) Impact evaluation: A guide for commissioners and managers. Report, Bond, London, May.
Thiem A (2014) Navigating the complexities of qualitative comparative analysis: Case numbers, necessity relations, and model ambiguities. *Evaluation Review* 38(6): 487–513.
Thiem A (2017) Conducting configurational comparative research with qualitative comparative analysis: A hands-on tutorial for applied evaluation scholars and practitioners. *American Journal of Evaluation* 38(3): 420–33.
Tipton E (2013) Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation Review* 37(2): 109–39.
Uprichard E (2009) Introducing cluster analysis: What can it teach us about the case? In: Byrne D and Ragin CC (eds) *The SAGE Handbook of Case-Based Methods*. Thousand Oaks, CA: SAGE, 132–47.
Van Draanen J (2016) Introducing reflexivity to evaluation practice: An in-depth case study. *American Journal of Evaluation* 38(3): 360–75.
Van Voorst S (2017) Evaluation capacity in the European Commission. *Evaluation* 23(1): 24–41.
Vellema S, Ton G, De Roo N, et al. (2013) Value chains, partnerships and development: Using case studies to refine programme theories. *Evaluation* 19(3): 304–20.
Verweij S and Gerrits LM (2013) Understanding and researching complexity with Qualitative Comparative Analysis: Evaluating transportation infrastructure projects. *Evaluation* 19(1): 40–55.
Wilkinson H, Hills D, Penn AS, et al. (2021) Building a system-based theory of change using participatory systems mapping. *Evaluation* 27(1): 80–101.
Yin RK (1997) Case study evaluations: A decade of progress? *New Directions for Evaluation* 1997: 69–78.
Yin RK (2013) Validity and generalization in future case study evaluations. *Evaluation* 19(3): 321–32.

Corey Schimpf is an Assistant Professor in the Department of Engineering Education at University at Buffalo. His work focuses on enhancing research through computational tools and advancing applied research methods.

Pete Barbrook-Johnson is a Senior Research Fellow in the Department of Sociology at the University of Surrey and a member of the Centre for the Evaluation of Complexity Across the Nexus (CECAN).

Brian Castellani is a Professor of Sociology at Durham University. His research involves advancing the tools of social complexity theory and computational social science for public health and policy research.