CAMBRIDGE Mathematics

Evaluating the Cambridge Mathematics Framework

Perspectives and approaches

Authors

Ellen Jameson

Representing the work of

Lynn Fortin, Tabitha Gould, Rachael Horsman, Ellen Jameson, Vinay Kathotia, Ray Knight, Lynne McClure, Darren Macey, Dominika Majewska, Nicky Rushton, Lucy Rycroft-Smith and Ben Stevens.

CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT

UNIVERSITY OF CAMBRIDGE Faculty of Mathematics

UNIVERSITY OF CAMBRIDGE Faculty of Education

CAMBRIDGE
√Mathematics

Evaluating the Cambridge Mathematics Framework: Perspectives and approaches

Sections in this document

CAMBRIDGE UNIVERSITY PRESS & ASSESSMENT
UNIVERSITY OF CAMBRIDGE Faculty of Mathematics
UNIVERSITY OF CAMBRIDGE Faculty of Education

Introduction: Why evaluate

The difficulty of determining the impact of one design influence among many makes evaluation of educational designs a nuanced undertaking. When a tool for educational design is to be evaluated, the increasing distance between the design tool, the finished design, the teacher and the student adds an extra layer of complexity to the evaluation process. Evaluation must therefore involve multiple tailored approaches which acknowledge and illuminate the interplay of contexts and factors involved. The nature of design makes this level of detail and variation necessary: "the economic and societal necessity for continuous improvement in education dictates that researchers and reformers engage in the design of tools, environments, and systems without knowing beforehand…all of the relevant parameters that impact their eventual success" (Middleton et al., 2006, p. 8).

An evaluation framework can help us to justify and structure the full range of research objectives so that individual studies, both quantitative and qualitative, appropriately characterise, contextualise and communicate the contributions of a design tool. This can make the implications of multiple studies more transparent and useful for stakeholder decision-making. Evaluations which take the project's theoretical influences into account may also contribute to future design efforts and continuing development of the underlying theories. This paper presents a framework to structure evaluation of the Cambridge Mathematics Framework (CMF), a tool for educational design in mathematics, and makes the case for the importance of relying on diverse indicators and research designs when interpreting outcomes which result from use of the CMF.

Evaluation goals and intervention studies

For the purposes of evaluation research, an intervention is a deliberate change to something that is being done in a real-world setting with the goal of achieving desired results (outcomes) that might not have been likely or perhaps even possible otherwise. Some outcomes of an intervention may be concrete and straightforward, while others might be more difficult to trace, whether because they cannot be measured directly, because they occur "downstream" from the direct intervention (secondary effects) or because other influences at work in a particular context make it hard to attribute outcomes to specific effects of an intervention (Middleton et al., 2006; Stern, 2015). In general, the goal of project outcome evaluation is to establish causal links between interventions and outcomes which are sufficient to justify decisions about whether it is worth continuing, expanding or copying aspects of the intervention (Schmitt & Beach, 2015; Stern, 2015; Stern et al., 2012).

Due to the complexity of forces at work in many real-world situations, a mix of direct and indirect outcomes from an intervention is common. In order to make sense of multiple types of outcomes, evaluations would ideally involve multiple intervention studies, each of which is tailored to particular aspects and contexts of the interventions (Delahais & Toulemonde, 2012; Stern et al., 2012). For effects which cannot be measured directly, it is also necessary to choose and justify proxy indicators whose distance from the effect itself must be taken into account when results are interpreted.

An evaluation framework is used to organise evaluation objectives, evaluation questions, indicators, success criteria, study contexts and methods so that studies can be designed to yield meaningful results and conclusions can be drawn accordingly. For interventions designed according to principles derived from theoretical influences, an evaluation framework can also be linked to the logic model indicating the reasoning behind how a design is expected to work, and results from evaluation studies can be used to further develop underlying theories and design principles (Barab, 2014; Cobb et al., 2003).

Impact models help to position the range of direct and indirect outcomes within the bigger picture of the overall impact these outcomes are expected to have in the world. In order to understand why this is useful, it helps to clarify the difference between outcome and impact. While these terms are sometimes used interchangeably, in evaluation the term *impact* generally refers to meeting the broadest goals of the project, while *outcomes* can (somewhat) more easily be identified as resulting from specific interventions. Outcome evaluation is therefore narrower in scope than impact evaluation would be. However, in both cases it is important to keep impact models in mind in order to decide on appropriate interventions and study designs.

Approaches to theory-based evaluation in education contexts

Causality is commonly a weak point in evaluation (Mayne, 2012), and this is not surprising given the range of data and depth of logic models that may be necessary to support it (Delahais & Toulemonde, 2012). Nevertheless, the complexities of real-world education settings make an understanding of what is happening, and why, essential for good decision-making. The term *theory-based evaluation* can describe any approach which seeks not only to determine whether a desired outcome is reached or not but to explain why, given relevant details of the context and the design being implemented (Mayne, 2012; Schmitt & Beach, 2015). Educational interventions involve many external factors which cannot be controlled, whether because of ethical issues surrounding interventions in students' and teachers'

classroom experiences (Burner, 2016) or because it would reduce the authenticity of the intervention context and thus render unclear the implications for more authentic contexts or for theories involving them (Barab, 2014). Therefore the ability to trace back from the outcomes, through the circumstances which produced the data, to the contributions of factors of interest is essential for interpreting outcomes in most education contexts.

The use of logic models to keep track of the features of a design which have the opportunity to influence the outcomes of an intervention supports the exploration of causality. This approach is common to many forms of theory-based evaluation (Beach & Pedersen, 2013; Delahais & Toulemonde, 2012; Schmitt & Beach, 2015). In education, design-based research (DBR) practices typically involve some form of logic model linking theories influencing design to features, actions and outcomes, so that outcomes can contribute to theory development (Barab, 2014; Cobb et al., 2003; diSessa & Cobb, 2004). Such models may also be created as part of educational design research practices more broadly (McKenney & Reeves, 2012; van den Akker et al., 2006). Likewise, conjecture mapping is a technique for mapping across logic model elements which is sometimes used to articulate relationships between different theories, design components and outcomes more explicitly, and to highlight potential gaps in theorising or design features (Sandoval, 2014).

Contribution analysis (CA) is a method in program evaluation with some fundamental similarities to DBR, including a focus on tracing the path through theory, design, implementation and outcomes (Delahais & Toulemonde, 2012; Mayne, 2012), and accounting for alternative explanations for observed outcomes (Lemire et al., 2012). However, in contrast to DBR, a primary goal of CA is then to look across implementations systematically, bringing many forms of data together from different contexts to tell a contribution story: a more general narrative explaining what seems to succeed when and why (Delahais & Toulemonde, 2012). Also in contrast to DBR, the main purpose of this story is to inform decision-making about adoption of a program which has been designed. We intend evaluation of the CMF to inform this type of decision-making as well, and so we combine some of the theoretical strengths of approaches from DBR and conjecture-mapping in education with CA to structure our evaluation framework.

Stern et al. (2012) note that more traditional approaches to evaluation, in which "causality is established by seeking a strong association between a single cause and a single effect…," are "not usually able to untangle the complexities of causal relations when causes are interdependent and affect outcomes as 'causal packages' rather than independently" (p. 38). An analysis of interventions in international development concluded that most interventions succeed (or fail) intertwined with other factors "as part of a causal package" where the outcomes are due to a combination of factors acting together

(Stern et al., 2012, p. 36). Although a comparable review of educational interventions with the same CA evaluation focus seemingly has not been undertaken, educational intervention studies commonly note factors outside the scope of research which may have influenced the outcomes in important ways. The concept, if not the name, of a causal package is familiar in the curriculum space, where student performance outcomes depend not only on the curriculum itself but how other designers and teachers enact it (Remillard & Heck, 2014). It is doubly applicable to a design tool which is another step removed from student outcomes.

Given the chains of direct and indirect outcomes we anticipate from the use of the CMF, causality is an important focus for our evaluation efforts. Our evaluation framework is designed to support this focus by structuring contribution analysis at the level of each study and across studies. While the usefulness of CA can be limited if theories of contribution and causality are overly complex (Delahais & Toulemonde, 2012) or overly simple (Lemire et al., 2012), this limitation can be overcome by determining sufficient levels of detail (Downes et al., 2019) based on theoretical foundations and past experience. Our prior formative case study research has allowed us to develop logic models at a feasible level of detail, which we can apply not only to research design but to interpretation of results. As this approach is applied across a number of studies, it will enable us to tell the contribution story of the CMF within the wider landscape of impacts.

Approaches to assessing the strength of evidence in an evaluation

Intervention studies are designed to address the interests and perspectives of different stakeholders. While randomised controlled trials (RCTs) are considered rigorous in medicine and throughout the social sciences, including in education, they are not always considered the most appropriate way to approach complex research contexts with uncontrolled factors, like classrooms. Outhwaite et al. (2020) note that "RCTs are argued to be too reductionist for evaluation studies conducted in complex environments, such as schools (Biesta 2010). Specifically, the emphasis on statistical aggregation removes educational interventions and their outcomes from their situated context (Elliott 2001)" (p. 225). This should not mean that an RCT could not provide useful evidence for us, but it does suggest that other study designs may be as, or more, relevant, particularly with the causal focus we have adopted.

The educational evaluation community as a whole offers some general guidance on the quality of evidence provided by different study designs for different purposes. In the UK, the Evidence for Policy and Practice Information and Coordinating Centre (EPPI-Centre) has developed the Weight of Evidence

Framework for rating the quality of evidence from different kinds of studies. This framework supports judgement in three dimensions: (a) internal validity, (b) appropriateness of study method for addressing the review question (in our case, evaluation questions) and (c) appropriateness of samples, context and measures. Each of these three dimensions are added to produce a combined weight indicating "the extent to which the study contributes evidence" (Gough, 2007, p. 11). This approach offers some flexibility in the use of specific evidence evaluation frameworks, and EPPI-C teams report applying it differently, with some prioritising the third dimension, some the first, some making no distinction between RCTs, quasi-experimental and non-controlled trials, others making that distinction – but with most weighting experimental evidence the highest (Gough, 2007).

Similarly, the US Department of Education's Institute for Education Sciences (IES) has developed the What Works Clearinghouse (WWC), a system for reviewing and reporting research on "programs, products, practices, and policies" (IES, 2021). The WWC Procedures and Standards handbooks provide a review framework for evaluating educational research to provide perspectives on whether and how evidence should contribute to decision-making or practice. The WWC framework puts studies of groups (e.g. of students, teachers, designers, schools) into three categories: RCTs, quasi-experimental design (QEDs) and regression discontinuity design (RDDs). It contrasts these with single case design (SCD) studies which might focus on a single school's scheme of work, a single designer or teacher's practices or knowledge, a single student, etc. Within each category, further guidance is provided for what constitutes good research design and appropriate treatment of data and claims (What Works Clearinghouse, 2017). In providing the results of reviews to an audience of education professionals, the WWC does not focus on the hierarchy of evidence from different study designs, Rather, a review rates whether a study meets WWC standards or not relative to the type of study it is (IES, 2021).

Notably, both databases distinguish between quality of evidence and strength of evidence, and neither includes comparative evidence hierarchies in their quality review protocols. Quality of evidence is relative to the study design, whereas strength of evidence is relative to how the evidence should be used by decision-makers. For example, the WWC adds labels to studies according to an external evidence hierarchy mandated by the Every Student Succeeds Act (ESSA), a policy designed to encourage schools to adopt evidence-based interventions. This hierarchy, reflecting the evidence valued by the policy-makers for this purpose, places good RCTs at the top, followed by good QEDs, good correlational studies and studies which "demonstrate a rationale" (*Evidence-Based Interventions Under the ESSA - Every Student Succeeds Act (CA Dept of Education)*, 2021).

Other generalised evidence hierarchies in the social sciences tend to rank well-powered large RCTs the highest, followed by underpowered RCTs, non-randomised experimental studies, non-randomised studies with historical controls, and case studies (Nutley et al., 2013). The norms reflected in these hierarchies are worth noting, but should not be assumed to match the norms of our stakeholder groups and do not determine which research questions and study designs are most relevant for our evaluation. For example, our evaluation's causal focus places values on the high internal validity of case studies.

As we design intervention studies for this evaluation we will consider EEPI-C and WWC guidelines along with other sources in developing study designs. The contribution analysis approach itself is neutral with respect to quality criteria and evidence hierarchies; evaluators must determine what criteria are relevant, and what standards should be met, in particular contexts (Klaver et al., 2016). The types of study designs which are most available to us at the moment, based on current opportunities and resources, are single case studies and case study comparisons. As we progress we may undertake case-control studies, QEDs, RCTs and even longitudinal studies. Each type of study in the appropriate context could contribute something useful to evaluation of different aspects of the nature of the CMF.

Approach to evaluating the CMF

Interventions involving the CMF tool and techniques for using it include design and implementation of curricula, resources, professional development (PD) materials and assessments in mathematics education according to principles of coherence, connectedness and professional communication.[1] Our evaluation framework will help us to structure multiple evaluation objectives for implementations of the CMF, aligned with our impact model. The nature of the CMF as an educational design tool, the impact model informing our design, and our evaluation objectives provide the background for the evaluation framework itself. The sections which follow will provide background for the evaluation framework in two parts:

A. The nature of the CMF as an educational design tool

B. The impact model informing our design

Both of these will then be linked to the description of the evaluation framework.

[1] These are summarised in Fig. 2 below. More detail can be found in *A Manifesto for Cambridge Mathematics* (McClure, 2015) and *An update on the Cambridge Mathematics Framework* (Cambridge Mathematics, 2018)

A) The CMF as an educational design tool

The potential scope of influence of a design tool is large, encompassing designers' outputs within an education system and teaching and learning outcomes when designs are implemented. For our purposes an *educational design* is anything intentionally developed to contribute to positive teaching and learning outcomes, according to designers' beliefs about what would successfully make this contribution and how. It might be as large as a curriculum, as small as a classroom task or a single day's lesson, or anything in between (textbooks, software, videos, etc.). The *educational design process* is the system of designers' professional activities resulting in a completed or refined design. *Educational designers* might be anyone engaged in educational design with a role in one or more of the multiple professional communities which can be found working at different scales: current or former classroom teachers, subject specialists, curriculum committee members, educational researchers, and others.

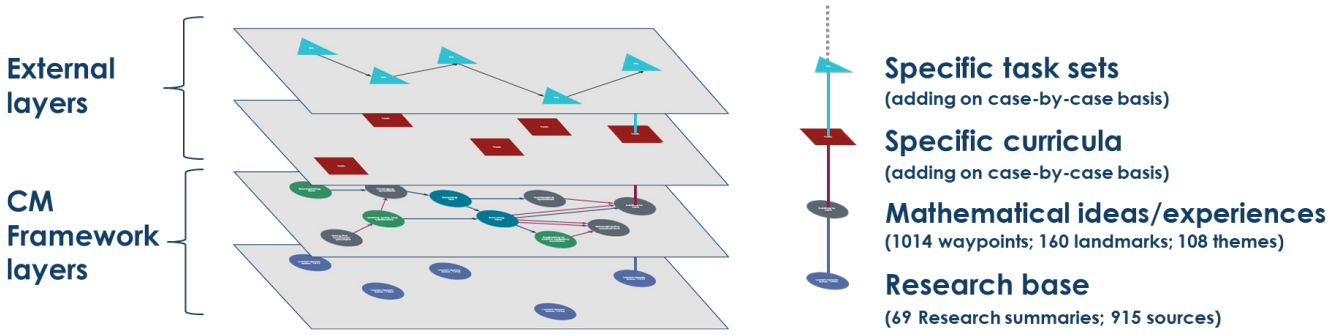
An *educational design tool* provides support for educational design processes, which may include making decisions about educational content and presentation, alignment with implementation contexts and facilitating design discussions among and between designers and stakeholders.

The CMF is an educational design tool which presents a searchable network of interdependent ideas in school mathematics which designers can use for reference and analysis as they make design decisions (see Figure 1). This network is

- derived from interpretation and synthesis of research and mathematics education,
- linked to the underlying research base and to specific curricula or task libraries, and
- expressed with multiple forms of supporting documentation accessible to users with different backgrounds and varying degrees of classroom experience.

Figure 1 on next page

Figure 1: Diagram showing the connected layers of the CMF



Multiple aspects of the CMF are relevant to consider for evaluation. It may be thought of as

- a decision aid,
- a conceptual model,
- an information system serving dynamic knowledge maps, and
- a shared frame of reference for research implications in design.

Each of these perspectives suggests particular ways in which we might evaluate outcomes from the use of the CMF to inform further refinement of both the design of the tool and our impact model.

A decision aid

The purpose of decision aids is to help choices to be made for good reasons while minimising irrelevant information. Decision aids allow the user to extend their range of experience with additional information so that their decisions are as informed as possible. They should incorporate practical considerations, and be made with as much knowledge as possible of the likely value of possible outcomes. The use of decision aids can make stakeholders more satisfied with the outcomes, particularly if they are represented in shared decision-making, though results from using decision aids vary with circumstances and they do not guarantee desired outcomes (O'Connor et al., 1999). A decision aid should be acceptable to target audiences, decrease uncertainty and conflict around making choices and help users to apply knowledge effectively (Nelson et al., 2007). This lends credence to the agency such tools leave in the hands of decision-makers; the decisions which result may be influenced by the rest of their professional experience as much or more as the decision aid.

Our most immediate goal for the CMF is that it should help its users to keep aware of important connections and dependencies between mathematical ideas as they make decisions about the

inclusion, organisation, alignment and communication of educational designs. It should also facilitate shared decision-making about design, among and between designers and stakeholders. We believe this awareness of connectivity in design and communication may support better and more productive experiences with mathematics for students and teachers (Cambridge Mathematics, 2018; Jameson et al., 2018). These decisions may be about a national-level curriculum, a school-level curriculum or scheme of work, a series of textbooks, a unit or lesson plan, or a single classroom activity, and are often taken by someone with some classroom teaching experience and/or some experience with mathematics as a subject.

A conceptual model

The network of mathematical ideas in the CMF is a conceptual model. Such models are used for simplification (Barlas & Carpenter, 1990) and visualisation of important structures and relationships (Seel, 2004). They can also represent "shared knowledge of a discipline" (Seel, 2004, p. 56) to members of different communities. A conceptual modelling approach has allowed us to highlight ideas and connections central to existing research on mathematics learning. The CMF map interface allows users to filter and visualise areas of the model as network maps, with additional detail available for features in the map, and narrative documents explaining the map as a whole. As is the case for other conceptual models, the CMF model has been informed by research on existing relevant models (Seel, 2004): individual mental models of mathematics (students, teachers and other educational designers), design and instructional models (teachers and educational designers), and psychological models of mathematics learning (held by researchers and CMF designers).

Our goals for the CMF as a conceptual model are that it should be trustworthy, meaningful and useful – that is, it should reflect a valid interpretation of the implications of an appropriate selection of research, and that this should be relevant to and accessible for design and teaching decisions. Our initial evaluation of the the conceptual model has been through expert review of conceptual descriptions and connections and how these have been interpreted from the literature in specific content areas (Jameson, 2019).

An information system for dynamic knowledge maps

Output from the CMF (data and relationship mappings) can be expressed as dynamically generated knowledge maps, visual representations of key ideas and relationships as the nodes[2] and edges[3] of a graph. The affordances and problems of knowledge maps as reference tools have been generally characterised through studies of knowledge representation in education (Stahl, 2006) and knowledge

[2] mathematical ideas in the CMF, represented as points in a map
[3] relationships between two mathematical ideas in the CMF, represented as lines connecting them in a map

management (Eppler, 2004). In general, good maps can address these questions: Where am I in the landscape, where can I get to from here, what route do I want to take, and what resources will I need for the journey? Good knowledge maps also need to address the questions, "how do I find relevant knowledge, how can I judge its quality, how can I make sense of its structure, and how do I go about applying it or developing it myself?" (Eppler, 2004, p. 192).

This knowledge mapping approach facilitates what Moss (2013) calls knowledge flows between research and practice in education, because (a) subject specialists with teaching as part of their backgrounds synthesise research to create the maps, (b) researchers review them, and (c) researchers, designers and teachers access them for various purposes. Evaluation is a further step towards communicating and improving the support for knowledge flows in the maps the CMF currently provides.

A shared frame of reference for conceptual connections and research implications

Members of different communities in mathematics education may each have distinct aims and backgrounds. Consequently they may find different aspects of connected mathematical ideas meaningful – what they look like for students in the classroom, how they can depend on one another, what research has influenced their inclusion. Explicit and shareable representations of mutually relevant information can facilitate coordination of perspectives and work between professional communities (Lee, 2007; Star & Griesemer, 1989), i.e. having a shared frame of reference may support the flow of knowledge between them. In the CMF, research reports, maps, descriptive text, images and classroom examples are twined together so that those in different roles can recognise and understand the information they need, and can gain insight into how a mathematical idea is approached from other perspectives which may impact their work. For example, an educational designer may need insight into how and when a teacher might find it appropriate to use a particular activity with their students.

The distinction between the CMF and a curriculum

There are key differences between the CMF and a curriculum which have implications for evaluation. In curriculum evaluation, student performance data is commonly used as one indicator of curriculum quality. It would be tempting to assume that classroom outcomes could be used as the same kind of indicator for the CMF as they are for a curriculum; however, while still useful, this type of indicator cannot be interpreted in the same way. A curriculum is already at least one step removed from the classroom; it goes through a process of translation when it is enacted which can result in a range of different classroom experiences. As a design tool, the CMF is two steps removed from the classroom; it does not prescribe a single selection or sequence of mathematical ideas but offers informed choices for designers

to make according to their circumstances. This means that student performance data may be even more strongly influenced by other factors. It could be important to include student performance data as an indicator, but it would be equally important to be clear about what it indicates.

Table 1: Comparing the CMF design tool to a curriculum

Cambridge Mathematics Framework	Curriculum
Structures ideas in a network according to conceptual interdependencies	Structures ideas in a linear sequence according to key objectives
Contains multiple alternate paths between two ideas which do not necessarily imply teaching order	Contains one path (per track)
Contains as many key ideas as can be synthesised from the literature	Contains a selection of key ideas shaped by constraints of the curriculum setting
Designed to be flexibly applicable across different educational contexts	Designed to be specifically applicable to a specific educational context
Describes mathematical ideas as initial, intermediate and culminating experiences	Describes resulting performance objectives

B) Impact models for the CMF

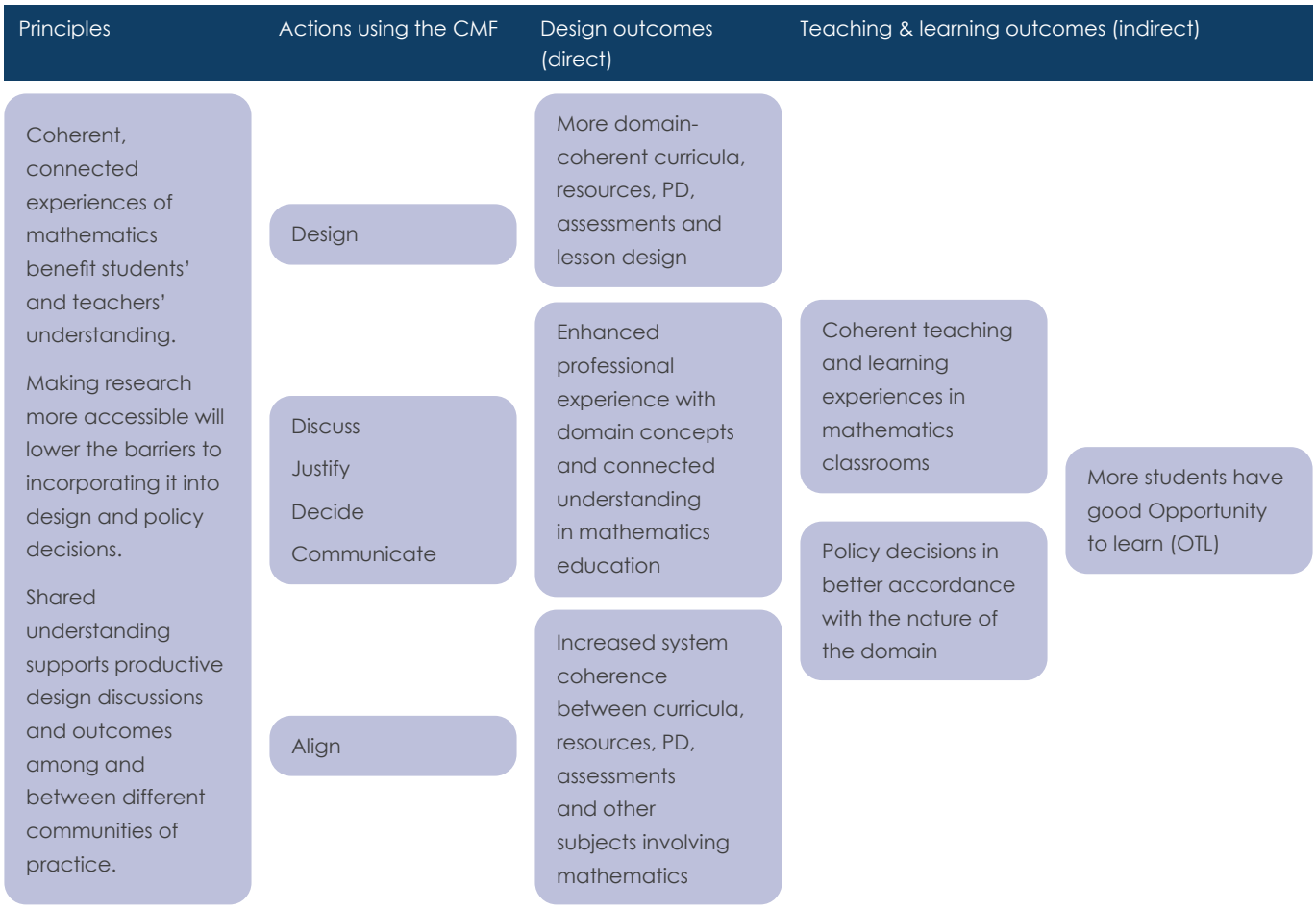
We use a high-level impact model to inform our high-level goals for evaluation and to set priorities when planning evaluation studies. We use detailed impact models, which are fine-grained and context-specific, when designing individual studies. Together these sets of impact models provide a complete picture of mechanisms for impact.

High-level impact model

We have described our goals for the CMF as part of the wider Cambridge Mathematics agenda in our manifesto and subsequent updates (Cambridge Mathematics, 2018; McClure, 2015). Figure 2 provides a summary of our high-level impact model, highlighting the three principles which have most strongly shaped the design of the CMF and linking these to the actions describing how the CMF is used. In turn we expect these actions to lead to certain direct types of design outcomes, which themselves would be expected to lead to indirect outcomes. Taken together both types of outcomes comprise the intended impacts of the CMF. We also believe that students' increased opportunity to learn mathematics through coherent and connected learning experiences in the classroom will improve the ways in which they

use maths in all areas of their adult lives, but impact at that level of remove is outside the scope of this evaluation framework.

Figure 2: High-level impact model



Key definitions

We have developed specific definitions for the key concepts, actions and actors in our model:

- *Domain coherence* refers to how designers make use of the connected structure of ideas which build on one another in the mathematics learning domain (Michener, 1978; Tall, 2013; Thurston, 1990).
- *System coherence* refers to conceptual and temporal alignment of curricula, assessments, resources and teacher professional development (Schmidt et al., 2005).[4]

[4] Jameson et al. (2018) describe our perspectives on domain coherence and system coherence in detail.

- *Shared understanding between communities of practice* refers to the priorities and perspectives on conceptual connections which can differ between the various professional communities in mathematics education.
- *Curriculum* is designated knowledge to be taught and learned, organised in time as discrete chunks of content (Bernstein, 1971; Young, 1971), which may span multiple years or less than a year of study and be documented in varying amounts of detail depending on whether it is intended for broad guidance or for detailed planning at the school level.
- *Resources* are materials teachers can access to use with students in lessons, including textbooks, activities and tasks.
- *Teacher professional development* refers to resources and programmes which help teachers to develop their professional knowledge.
- When we refer to *assessments*, we are often referring to assessment frameworks which structure standardised summative assessment of student performance, used as an indicator of the received curriculum (Remillard & Heck, 2014). We believe that summative and formative classroom assessments can also benefit from connected, coherent design.

Detailed impact models

Detailed impact models describe what is observable, what is measurable, and what is potentially attributable to the influence of the CMF at the interaction level in a given scenario. Knowing this helps us to develop specific research questions and methods for evaluation studies. For direct design outcomes, our impact models describe how designers' interactions with various features of the CMF are expected to produce the desired design outcomes. For indirect teaching and learning outcomes, our detailed impact models describe how interactions with designs which have been influenced by the CMF are expected to lead to particular outcomes.

To develop these models, we use conjecture mapping methods to create logic models for specific uses of the CMF, with impacts as endpoints. Several case-specific models at this detailed level have been produced[5] and we are beginning to generalise models across cases as we continue to collect more data. In order to do this it is important to define the range of circumstances to which each applies; we discuss this further in the Distance range component of the evaluation framework.

[5] E.g. Jameson & Horsman, 2019, 2020; Majewska, 2021

Evaluation Framework

We have structured the evaluation framework into eight components. Each component contributes to making intervention-specific decisions about study design, site selection and analysis. Taken together, these components also help to position data from multiple completed interventions within the larger contribution story which will present and explain outcomes from CMF use.

1. Evaluation goals

Our aims for evaluation are (1) to provide information that users and stakeholders need to know about the degree to which the CMF is a trustworthy, meaningful and useful tool in their contexts, and (2) to refine aspects of the CMF and our impact models, contributing to theory-building when appropriate.

Our evaluation goals fall into four categories, adapted from Stern et al. (2012) and Befani & Mayne (2014):

1. Attribution: To what degree are the outcomes due to influence of the CMF?
2. Contribution: Did the intervention have the outcomes that users of the CMF desired?
3. Mechanism: How did use of the CMF contribute to the outcomes?
4. Translation: Looking across studies, under what conditions would we expect to achieve the desired outcomes?

Any of these goal types could be relevant to any of our impact contribution claims (see examples of these in Table 2). The first two types of goals yield information about the existing relevance and quality of the CMF approach. The third and fourth involve developing an understanding of what contexts and conditions are appropriate for successful use of the CMF. We will need to design evaluation studies differently to address these goals depending on whether specific research questions we identify require data on direct or indirect outcomes of the use of the CMF. For each intervention, we will need to know in detail how the CMF is being used and what is happening in the range of outcomes which result in order to (a) verify that the CMF is an influence in the intervention, and if possible how central an influence it is, (b) interpret the outcomes accordingly and (c) use the data to strengthen the case for specific features and topic areas of the CMF or inform any changes that may be warranted.

We have chosen to model our approach on contribution analysis instead of realist evaluation so that we can focus on the multi-dimensionality we believe will be needed to explain what is happening and how. Even so, for several reasons we will approach our four types of evaluation objectives from a realist perspective overall (Westhorp et al., 2011). First, our casting of the CMF as a decision tool frames our belief that impact is caused by the people using the CMF to make decisions, though we also believe that the CMF is never provably the only factor in their decision-making. Second, the wide variety of contexts in which the CMF is intended to be useful means that CMF use will interact with a variety of other factors, and multiple types of success criteria, standards and priorities must be considered. Third, throughout the design of the CMF and synthesis of content we have viewed stakeholder participation as an essential means of considering the underlying construct validity and ecological validity of the tool (Jameson, 2019). Fourth, we intend to keep adding to our pool of evaluation data, and we expect to improve our understanding of uses and impacts of the CMF over time.

Evaluation goals and research questions specific to each study will be developed from each goal category listed above. A bigger picture will emerge of the range of uses and outcomes of the CMF as more intervention studies are carried out. Table 2 provides examples of research questions which might be appropriate for interventions focusing on each of the categories of impact contribution claims defined with respect to the high-level impact model (p. 11).

Table 2: Example research questions

Example CMF contribution claim	Example research questions
Contributions to classroom outcomes	What evidence shows that the CMF had a substantial influence on resources or teaching sequences underlying observed outcomes? (Goal 1, 3) Are there indications that teachers' comprehension of mathematical ideas have changed? Is this reflected in their teaching? (Goal 2) Are there indicators that students' mathematical behaviour has changed (in comparison with a matched or historical control)? In what way have their actions changed? (Goal 2) Looking across case studies in which the CMF was shown to have contributed, are there contextual characteristics which correlate with different types of outcomes? (Goal 4)

Example CMF contribution claim	Example research questions
Contributions to domain coherence	Is the CMF presenting a coherent picture to users depending on their queries? Does it improve their use of important connections? (Goal 2) Are the curricula/materials used in this study more coherent as a result of influence by the CMF? (Goal 1) Do their users develop a coherent or more coherent use of important connections which can be attributed in part to their interaction with these curricula or materials? (Goal 1, 2, 3)
Contributions to system coherence	Does the CMF make it feasible/easier to analyse and improve the alignment of curricula, resources, PD support and assessments with the corresponding curriculum? (Goal 1, 2, 3)
Contributions to professional decision-making	Do direct users of the CMF feel they have made more informed decisions? Do they feel this has improved the quality, efficiency or defensibility of their outputs? (Goal 2) Are these feelings borne out by downstream indicators of the quality of their design outputs? (Goal 1, 3)
Contributions from a map-based information system	Is useful information available? Can users access it easily? Is it expressed meaningfully and clearly for target audiences? (Goal 2) Do users find the CMF meets Eppler's (2004) criteria for good knowledge maps? (Goal 2)

2. Temporal range

We define the temporal range of a CMF intervention study as the amount of time over which the study must take place in order for relevant data to be available. It is likely to be positively correlated with the conceptual, systemic and distance ranges described below. On the shorter end, a study looking only at direct outcomes for designers using the CMF in a limited way might yield data about designer-level behaviours and outcomes after a few weeks. On the longer end, a study focusing on our system coherence goals might involve changes in curriculum alignment that teachers and students must experience over many years before enough of the required data can be collected. When a short intervention involves student outcomes (on the order of a few lessons-worth of study), effect sizes of any kind are likely to be small relative to the effects that could result from long-term interventions, but are more likely to be clearly detected and attributable to the intervention (Kraft, 2020).

3. Conceptual range

The conceptual range of a CMF intervention study is the number of mathematical ideas and relationships in the CMF which designers (and possibly then teachers and/or students downstream) would be expected to engage with. When a study focuses mainly on design activities, a larger conceptual range may require more time for designers to work with, in the order of weeks or months depending on the scale of the design project. When school implementation of a design is involved, this conceptual range may have larger implications for the temporal range. For example, the process of revising an area of a mathematics curriculum may take months, but the area which has been revised may cover years of study in the school system.

4. Systemic range

Systemic range is an important consideration for studies involving our system coherence goals. Factors affecting the characteristics of an intervention might include how many elements of the system are involved in alignment using the CMF (curriculum, assessment framework, resources, teacher professional development, etc.), and to what extent and over what portions of the curriculum these are to be aligned. The scope of such interventions would then interact with temporal, conceptual and distance ranges as well. Potential elements of education systems in which alignment activities could have an impact will be identified with reference to Remillard and Heck's (2014) model of curriculum enactment and elaborations made as part of the ICMI 24 study to highlight systemic considerations for curriculum reform (Jameson & Bobis, in press).

5. Distance range

One of the most fundamental considerations when evaluating an educational design tool is the fact that the tool itself is several steps removed from what students experience in classrooms. With each step removed, more outside influences accompany the influence of the CMF, all of which may contribute to teaching and learning outcomes, perhaps even outweighing the influence of the CMF. We call these steps the distance the CMF is removed from the indicators. We likewise distinguish between direct outcomes which are at no distance and indirect outcomes which are some distance removed from the data collection context.

Our past evaluation efforts have been case studies of direct outcomes of CMF use. Future evaluation studies could involve indirect outcomes in classrooms based on lessons, resources or curricula. Such approaches would benefit from quantitative measurement of student performance but would also require accompanying qualitative data explaining the degree to which the CMF contributed to measured outcomes. Demonstrating such contributions would be important for the overall contribution story as it applies to our high-level impact model.

This distance between intervention and measurement is a known issue in educational impact evaluation. Ruiz-Primo et al. (2002) describe a parallel situation of distance between different forms of assessment and the curriculum as enacted in the classroom. They studied patterns of instructional sensitivity of assessments at different distances from the curriculum and found that while only small or no effects might show up in general standardised testing, effect sizes from forms of assessment which were successively closer to the enacted curriculum tended to increase (Ruiz-Primo et al., 2002), This pattern has since been confirmed in other intervention contexts (Hickey et al., 2009; Hickey & Zuiker, 2012).

We have adapted the assessment distance framework of Ruiz- Primo et al. (2002) to highlight the degrees of distance between the use of the CMF and different sources of data which are of interest to stakeholders (see Table 3). Our hope is that this will make it clear to stakeholders and researchers alike why different indicators are useful at different levels of distance, and why multiple levels of distance may be needed to interpret the contributions of the CMF to indirect outcomes (at the Close – Remote levels in Table 3). Depending on opportunities and research questions identified for particular interventions, it may be possible and desirable to collect data at multiple levels in order to strengthen the contribution story and the likelihood of recognising relevant effects when they occur.

Table 3: Examples of different levels of distance from direct use of the CMF; adapted from Ruiz-Primo et al. (2002)

	Immediate	Close	Proximal	Distal	Remote
Distance from CMF	Develop or use artifacts created directly from use of the CMF	Develop or use processes, outputs closely aligned with the immediate artifacts, content of CMF	Work with relevant knowledge, insights in CMF but not our pre-curated groupings of ideas	Design a resource or assessment based on a CMF-influenced curriculum	Demonstrate performance on general measures on equivalent topics

	Immediate	Close	Proximal	Distal	Remote
Distance from a specific use of the CMF	SoW itself + artifacts students and teachers produce in/for a single lesson	Activities related to but not mentioned by the SoW, teacher judgment	Ideas applied flexibly between specific topics in which students were first taught	Regional/state performance measures	National or international standardised assessments

6. Settings

Direct outcomes of CMF use occur in settings where educational designers are working directly with the CMF. These might be in curriculum committee meetings, publishing houses or other designer workplaces, in local school systems or individual schools. Indirect outcomes could occur in classroom teaching and learning settings or policy discussions. Table 4 provides examples of different settings which are relevant to different types of impact contribution claims that could be evaluated with respect to the high-level impact model (p. 11).

Table 4: Examples of contexts in which different claims could be appropriately evaluated

Example CMF contribution claim	Relevant contexts
Classroom outcomes	Classrooms Life skills
Domain coherence	Mathematics teaching & learning Teacher PD Maths curriculum design
System coherence/ curriculum coherence	Curriculum implementation and alignment strategies
Professional decision-making	Classrooms Curriculum bodies, assessment bodies, publishing houses, the output of private companies and individuals
Educational design outcomes	Curriculum bodies, assessment bodies, publishing houses, the output of private companies and individuals
Information system effectiveness	Point of direct use

7. Participants and stakeholders

Depending on the nature of the intervention, participants might include educational designers engaged in curriculum or resource design, teachers, students, teacher educators or policy makers. Stakeholder groups include the direct users of the CMF from the participant groups but also the organisations and individuals who invest time and resources in CMF design and evaluation (e.g. the University of Cambridge, our external reviewers and collaborators).

8. Outcome indicators and measures

Different stakeholder interests lend themselves to different perspectives on evaluation and the concept of quality. Melrose (1998), when discussing this issue in curriculum evaluation, describes three evaluation paradigms. These perspectives can be applied to a view of quality as fit-for-purpose (e.g. success criteria involve student performance outcomes and standards) or as transformation-by-participation (e.g. success criteria involve improvements in the agency and decision-making of the intended beneficiaries) – see Table 5.

Table 5: Conceptions of quality across evaluation paradigms applied to curriculum design (Melrose, 1998)

	Functional	Transactional (naturalistic)	Critical/emancipatory
Perspective	**Objective framing** • Does it efficiently produce knowledgeable students, skilled workers? • Have pre-set goals been met? • What are the problems with this curriculum? • Does it deliver on its mandate, funding requirements?	**Pluralistic/subjective framing** • How does this learning sequence/event appear to different stakeholders? • Should goals, processes be changed to better suit context or participants? • How could this be improved to promote better learning experiences? • How does the context of this curriculum affect learning?	**Power and structural framing** • What anxieties do students have about this test and how can we minimise them? • What's going on in theory and practice for you as a maths teacher? • Why are students dropping this class/subject? • Why do students love this class/subject?
Quality as fit-for-purpose	Predetermined standards and outcome thresholds	Negotiated standards and outcome thresholds	

Functional	Transactional (naturalistic)	Critical/emancipatory
Quality as transformation by participation	Participation in evaluation feeds in to decision-making + cyclical improvement	Participating teachers and students are empowered to identify + act on opportunities to bring about positive change

Depending on the goals of designers and stakeholders involved in each study, indicators of direct outcomes with designers as participants could involve

- characteristics of the design output (connectedness, content inclusion or sequencing compared to a previous version), the types of justifications designers provide to their stakeholders,
- perceived effectiveness of design conversations, or
- changes in designer professional knowledge (self-reported or directly assessed).

Indicators of indirect outcomes/impacts, when a design influenced by the CMF is implemented in classrooms, could involve

- student performance relative to a comparison group (pre-post assessments),
- teacher lesson delivery (video observation and debriefing interviews), or
- teacher lesson planning and use of resources and direct observation of teacher planning, diary self-reporting, interviews, or lesson plan analysis).

Once indicators have been identified, success criteria can be defined through literature review and consultation with stakeholders, and specific standards for these criteria can be explored.

Applying the evaluation framework

We have adapted a process for using this evaluation framework to study and report on the direct and indirect outcomes of CMF use in curriculum and resource design from an example reported by Klaver et al. (2016), with additional elements from Delahais & Toulemonde (2012) and Mayne (2012):

1. Work with stakeholders to identify what it looks like when impacts are achieved according to our goals for impact.
2. Work with stakeholders to identify and prioritise various outcomes of interest.
3. Identify and prioritise intervention opportunities on the basis of 1. and 2.
4. For each prioritised intervention:

   a. Develop research questions and identify potential sources of data with relevance to causal pathways (e.g. data on how/how much the CMF contributed to a design, whether that design was enacted as intended, etc.).
   b. Gather data.
   c. Analyse data and report quality of data relative to implementation research questions.
   d. Consider alternative explanations for observed outcomes and gather additional data to address these if necessary.
   e. Explain observed outcomes: how they compare to implementation goals and how they came about.
   f. Determine the contribution of the CMF and develop the local contribution story.
   g. Stakeholder review of contribution story and review of evidence.
   h. Respond to stakeholder review and refine local contribution story.
5. Periodically review data from multiple intervention studies, clustering data from different studies which apply to the same causal link in the impact models as appropriate.
6. Work with stakeholders to review the strength of available forms of evidence.
7. Develop the larger contribution story and present types and strength of evidence in a table.
8. Present evidence to stakeholders.

The transparency afforded by this evaluation process and the diversity of information it can provide about outcomes resulting from use of the CMF is intended to give various stakeholders with their own particular priorities the basis they need for judging whether the CMF should be used in their context. Stakeholder participation in judging both the appropriate data to collect and judgment of the extent of the contribution the CMF makes will be essential in both design and teaching and learning contexts, as there will always be aspects of design and classroom enactment with bearing on the contribution story to which researchers may not have access.

Example of evidence for a local contribution story

In a recent curriculum design case study, the Australian Curriculum, Assessment and Reporting Authority (ACARA) used the CMF as a design tool to inform revision of the Statistics and Probability strands of the Australian Curriculum as part of its Foundation – Year 10 review. Table 6 characterises this case study using the evaluation framework.

Table 6: Single case study example: ACARA Statistics and Probability revision

Evaluation framework dimensions	Example: ACARA Statistics and Probability revision
Goals: Evaluation questions	1. Was the ACARA team happy with the results of the curriculum revision? 2. To what degree can changes made to the Statistics and Probability strands be attributed to use of the CMF? 3. How was the CMF used and how did this contribute to the revisions made? 4. Are there additional contexts we would expect these processes and outcomes to apply to? 5. Are particular refinements to the CMF content or interface warranted based on feedback about these processes or outcomes?
Temporal range	Foundation – Year 10; in practice mainly secondary, small but important relevance to primary
Conceptual range	Statistics and probability
Systemic range	Designed curriculum, national level
Distance range	Immediate and proximal: focus on direct measures and design outcomes
Contexts/settings	National curriculum review: curriculum design team and reviewers
Participants, stakeholders & beneficiaries	Participants: ACARA review team: expert mathematics curriculum designers Stakeholders: ACARA curriculum review leaders, teachers, consultants Beneficiaries: teachers, students
Outcome indicators & measures	Single case study: design implementation Interview and diary data; revised curriculum; reviewer feedback; no student performance data

In this case, research question 2, "To what degree can changes made to the Statistics and Probability strands be attributed to use of the CMF?" would provide the foundation for demonstrating contributions of the CMF in the contribution story. Table 7 shows an example of an evidence analysis table which would be used to develop the contribution story for research question 2 from available evidence.

Table 7: Example evidence analysis table for the ACARA case study, research question 2 (after (Delahais & Toulemonde, 2012, p. 288)

Item of evidence	Source type	Confirming/ refuting	Causal mechanism	Strength of evidence*
ACARA designers reported specific curriculum changes which they attributed to insight from the CMF	Direct	Confirming	Intended contribution	Very strong
ACARA designers reported feeling more confident because they had research justification	Direct	Confirming	Intended contribution	Rather weak
Teachers reported excitement about the new curriculum	Indirect	Confirming	Intended contribution	Rather weak
The ACARA team reported that the CMF did not reduce time spent on design	Direct	Refuting	Other contribution	Very strong
The ACARA team reported that the CMF improved the quality of time spent on design	Direct	Confirming	Intended contribution	Very strong

From this example we could start to build a local contribution story:

> *All designers involved in the ACARA review have an initial programme of research to draw on which influenced their perspectives. However, the CMF made distinct contributions to design decision-making related to domain coherence, and a group of initial teacher reviewers responded positively to the changes.*

This local contribution story could be expanded to include data from the other research questions, and could be reported to stakeholders along with the details of the case and data sources. We could then use it along with other local contribution stories to inform the larger contribution story which relates to our large-scale impact model. Stakeholder feedback could inform further data collection, additional studies, and/or refinement of the local or high-level impact models.

*All evidence should be valid and relevant; strength of evidence here refers to its applicability to the contribution story for research question 2, on a scale ranging from Very Strong to Very Weak, which we can develop a rubric to support.

Conclusions

The evaluation framework encompasses a wide range of time frames, conceptual coverage, systemic scales, and indicators of impact, some of which will be much more feasible to address in the near term than others. Evaluation of the CMF is a long-term effort, the pace of which will continue to depend on the availability of necessary resources and opportunities for implementation. Some of the interventions which could contribute to our evaluation goals have already begun, while other types of interventions – especially those requiring long time frames and extensive commitments from school systems – may not take place without additional support. Results of studies will be reported internally as they occur and externally in periodic updates on the evaluation process.

References

Barab, S. (2014). Design-based research: A methodological toolkit for engineering change. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (2nd ed., pp. 151–170). Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.011

Barlas, Y., & Carpenter, S. (1990). Philosophical roots of model validation: Two paradigms. *System Dynamics Review, 6*(2), 148–166. https://doi.org/10.1002/sdr.4260060203

Beach, D., & Pedersen, R. B. (2013). *Process-tracing methods: Foundations and guidelines*. University of Michigan Press.

Befani, B., & Mayne, J. (2014). Process tracing and contribution analysis: A combined approach to generative causal inference for impact evaluation. *IDS Bulletin, 45*(6), 17–36. https://doi.org/10.1111/1759-5436.12110

Bernstein, B. (1971). On the classification and framing of educational knowledge. In M. F. D. Young (Ed.), *Knowledge and control: New directions for the sociology of education* (pp. 47–69). Collier-Macmillan Publishers.

Burner, T. (2016). Ethical dimensions when intervening in classroom research. *Problems of Education in the 21st Century, 73*, 18–26. https://doi.org/10.33225/pec/16.73.16

Cambridge Mathematics. (2018). *An update on the Cambridge Mathematics Framework*. Cambridge Mathematics. https://www.cambridgemaths.org/images/cambridge-mathematics-symposium-2018-framework-update.pdf

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13. https://doi.org/10.3102/0013189X032001009

Delahais, T., & Toulemonde, J. (2012). Applying contribution analysis: Lessons from five years of practice. *Evaluation, 18*(3), 281–293. https://doi.org/10.1177/1356389012450810

diSessa, A. A., & Cobb, P. (2004). Ontological innovation and the role of theory in design experiments. *Journal of the Learning Sciences, 13*(1), 77–103. https://doi.org/10.1207/s15327809jls1301_4

Downes, A., Novicki, E., & Howard, J. (2019). Using the contribution analysis approach to evaluate science impact: A case study of the National Institute for Occupational Safety and Health. *American Journal of Evaluation, 40*(2), 177–189. https://doi.org/10.1177/1098214018767046

Eppler, M. J. (2004). Making knowledge visible through knowledge maps: Concepts, elements, cases. In C. W. Holsapple (Ed.), *Handbook on Knowledge Management 1: Knowledge Matters* (pp. 189–205). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24746-3_10

Evidence-based interventions under the ESSA - Every Student Succeeds Act (California Department of Education). (2021). California Department of Education. https://www.cde.ca.gov/re/es/evidence.asp

Gough, D. (2007). Weight of evidence: A framework for the appraisal of the quality and relevance of evidence. *Research Papers in Education, 22*(2), 213–228. https://doi.org/10.1080/02671520701296189

Hickey, D. T., Ingram-Goble, A. A., & Jameson, E. M. (2009). Designing assessments and assessing designs in virtual educational environments. *Journal of Science Education and Technology, 18*(2), 187–208. https://doi.org/10.1007/s10956-008-9143-1

Hickey, D. T., & Zuiker, S. J. (2012). Multilevel assessment for discourse, understanding, and achievement. *Journal of the Learning Sciences, 21*(4), 522–582. https://doi.org/10.1080/10508406.2011.652320

IES. (2021). *WWC What Works Clearinghouse*. National Center for Educational Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/

Jameson, E. (2019). *Methodology: Formative evaluation*. Cambridge Mathematics. https://www.cambridgemaths.org/Images/methodology-formative-evaluation.pdf

Jameson, E., & Bobis, J. M. (in press). Modelling curriculum reform: A system of agents, processes and objects. In R. Vithal & Y. Shimizu (Eds.), *ICMI study 24: School mathematics curriculum reforms: Challenges, changes and opportunities*. Springer.

Jameson, E., & Horsman, R. (2019). *Writing a textbook chapter using the Cambridge Mathematics Framework* [Case study report]. Cambridge Mathematics. https://www.cambridgemaths.org/Images/writing-a-textbook-chapter.pdf

Jameson, E., & Horsman, R. (2020). *Using the Cambridge Mathematics Framework to refine the UNICEF-Cambridge Curriculum Progression Framework* [Case study report]. Cambridge Mathematics. https://www.cambridgemaths.org/Images/UNICEF-cambridge-curriculum-progression-framework.pdf

Jameson, E. M., McClure, L., & Gould, T. (2018). Shared perspectives on research in curriculum reform: Designing the Cambridge Mathematics Framework. In Y. Shimizu & R. Vithal (Eds.), *ICMI Study 24 conference proceedings* (pp. 531–538). ICMI.

Klaver, D., Mohapatra, B. P., & Smidt, H. (2016). *Enhancing confidence in contribution claims by lobby and advocacy programs* (No. EES16-0323). Centre for Development Innovation, Wageningen University and Research centre. https://edepot.wur.nl/399938

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Lee, C. P. (2007). Boundary negotiating artifacts: Unbinding the routine of boundary objects and embracing chaos in collaborative work. *Computer Supported Cooperative Work (CSCW)*, 16(3), 307–339. https://doi.org/10.1007/s10606-007-9044-5

Lemire, S. T., Nielsen, S. B., & Dybdal, L. (2012). Making contribution analysis work: A practical framework for handling influencing factors and alternative explanations. *Evaluation, 18*(3), 294–309. https://doi.org/10.1177/1356389012450654

Majewska, D. (2021). *Using the Cambridge Mathematics Framework to map the Common Core to HOTmaths* [Case study report]. Cambridge Mathematics. https://www.cambridgemaths.org/Images/hotmaths-case-study-a-summary.pdf

Mayne, J. (2012). Contribution analysis: Coming of age? *Evaluation, 18*(3), 270–280. https://doi.org/10.1177/1356389012451663

McClure, L. (2015). *A manifesto for Cambridge Mathematics*. Cambridge Mathematics. https://www.cambridgemaths.org/Images/cambridge-mathematics-manifesto.pdf

McKenney, S., & Reeves, T. C. (2012). *Conducting educational design research*. Routledge.

Melrose, M. (1998). Exploring paradigms of curriculum evaluation and concepts of quality. *Quality in Higher Education, 4*(1), 37–43. https://doi.org/10.1080/1353832980040105

Michener, E. R. (1978). *The structure of mathematical knowledge* (Technical Report No. 472). MIT Artificial Intelligence Laboratory. https://dspace.mit.edu/handle/1721.1/6928

Middleton, J., Gorard, S., Taylor, C., & Bannan-Ritland, B. (2006). *The "compleat" design experiment: From soup to nuts* [Research paper]. University of York, Department of Educational Studies. https://www.academia.edu/1913111/TheCompleatDesign_Experiment_from_soup_to_nuts

Moss, G. (2013). Research, policy and knowledge flows in education: What counts in knowledge mobilisation? *Contemporary Social Science, 8*(3), 237–248. https://doi.org/10.1080/21582041.2013.767466

Nelson, W. L., Han, P. K. J., Fagerlin, A., Stefanek, M., & Ubel, P. A. (2007). Rethinking the objectives of decision aids: A call for conceptual clarity. *Medical Decision Making, 27*(5), 609–618. https://doi.org/10.1177/0272989X07306780

Nutley, S., Powell, A., & Davies, H. (2013). *What counts as good evidence?* (p. 40) [Provocation paper for the Alliance for Useful Evidence]. Research Unit for Research Utilisation (RURU), University of St Andrews. https://www.alliance4usefulevidence.org/assets/What-Counts-as-Good-Evidence-WEB.pdf

O'Connor, A. M., Fiset, V., DeGrasse, C., Graham, I. D., Evans, W., Stacey, D., Laupacis, A., & Tugwell, P. (1999). Decision aids for patients considering options affecting cancer outcomes: Evidence of efficacy and policy implications. *JNCI Monographs, 1999*(25), 67–80. https://doi.org/10.1093/oxfordjournals.jncimonographs.a024212

Outhwaite, L. A., Gulliford, A., & Pitchford, N. J. (2020). A new methodological approach for evaluating the impact of educational intervention implementation on learning outcomes. *International Journal of Research & Method in Education, 43*(3), 225–242. https://doi.org/10.1080/1743727X.2019.1657081

Remillard, J. T., & Heck, D. J. (2014). Conceptualizing the curriculum enactment process in mathematics education. *ZDM, 46*(5), 705–718. https://doi.org/10.1007/s11858-014-0600-4

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39*(5), 369–393. https://doi.org/10.1002/tea.10027

Sandoval, W. A. (2014). Conjecture mapping: An approach to systematic educational design research. *Journal of the Learning Sciences, 23*(1), 18–36. https://doi.org/10.1080/10508406.2013.778204

Schmidt, W. H., Wang, H. C., & McKnight, C. C. (2005). Curriculum coherence: An examination of US mathematics and science content standards from an international perspective. Journal of *Curriculum Studies, 37*(5), 525–559. https://doi.org/10.1080/0022027042000294682

Schmitt, J., & Beach, D. (2015). The contribution of process tracing to theory-based evaluations of complex aid instruments. *Evaluation, 21*(4), 429–447. https://doi.org/10.1177/1356389015607739

Seel, N. M. (2004). Model-centered learning environments: Theory, instructional design, and effects. In N. M. Seel & S. Dijkstra (Eds.), *Curriculum, plans, and processes in instructional design* (pp. 49–73). Lawrence Erlbaum Associates.

Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge*. MIT Press.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, "translations" and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science, 19*(3), 387–420. https://doi.org/10.1177%2F030631289019003001

Stern, E. (2015). *Impact evaluation: A guide for commissioners and managers*. UK Department for International Development (DFID). https://assets.publishing.service.gov.uk/media/57a0896de5274a31e000009c/60899_Impact_Evaluation_Guide_0515.pdf

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). *Broadening the range of designs and methods for impact evaluations* [Technical report]. Institute for Development Studies. https://doi.org/10.22163/fteval.2012.100

Tall, D. (2013). *How humans learn to think mathematically: Exploring the three worlds of mathematics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139565202

Thurston, W. P. (1990). Mathematical education. *Notices of the AMS, 37*(7), 844–850. http://pi.math.cornell.edu/~mathclub/Media/thurston-mathematical-education.pdf

van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (Eds.). (2006). *Educational design research*. Routledge.

Westhorp, G., Prins, E., Kusters, C., Hultink, M., Guijt, I., & Brouwers, J. (2011). *Realist evaluation: An overview* [Seminar report]. Wageningen UR Centre for Development Innovation. https://core.ac.uk/download/pdf/29235281.pdf

What Works Clearinghouse. (2017). *What Works Clearinghouse Standards Handbook (Version 4.0)*. Institute of Education Sciences.

Young, M. F. D. (1971). An approach to the study of curricula as socially organized knowledge. In *Knowledge and control: New directions for the sociology of education* (pp. 19–46). Collier-Macmillan Publishers.