Rethinking programme evaluation in health professions education: beyond 'did it work?'

Faizal Haji,[1–3] Marie-Paule Morin[2,4] & Kathryn Parker[1,5]

CONTEXT For nearly 40 years, outcome-based models have dominated programme evaluation in health professions education. However, there is increasing recognition that these models cannot address the complexities of the health professions context and studies employing alternative evaluation approaches that are appearing in the literature. A similar paradigm shift occurred over 50 years ago in the broader discipline of programme evaluation. Understanding the development of contemporary paradigms within this field provides important insights to support the evolution of programme evaluation in the health professions.

METHODS In this discussion paper, we review the historical roots of programme evaluation as a discipline, demonstrating parallels with the dominant approach to evaluation in the health professions. In tracing the evolution of contemporary paradigms within this field, we demonstrate how their aim is not only to judge a programme's merit or worth, but also to generate information for curriculum designers seeking to adapt programmes to evolving contexts, and researchers seeking to generate knowledge to inform the work of others.

DISCUSSION From this evolution, we distil seven essential elements of educational programmes that should be evaluated to achieve the stated goals. Our formulation is not a prescriptive method for conducting programme evaluation; rather, we use these elements as a guide for the development of a holistic 'programme of evaluation' that involves multiple stakeholders, uses a combination of available models and methods, and occurs throughout the life of a programme. Thus, these elements provide a roadmap for the programme evaluation process, which allows evaluators to move beyond asking whether a programme worked, to establishing how it worked, why it worked and what else happened. By engaging in this process, evaluators will generate a sound understanding of the relationships among programmes, the contexts in which they operate, and the outcomes that result from them.

Medical Education 2013: ***47****: 342–351*
doi:10.1111/medu.12091

[1]Wilson Centre, University of Toronto, Toronto, Ontario, Canada
[2]SickKids Learning Institute, Hospital for Sick Children, Toronto, Ontario, Canada
[3]Department of Clinical Neurological Sciences, London Health Sciences Centre, London, Ontario, Canada
[4]Division of Immunology and Rheumatology, Department of Paediatrics, Saint-Justine Hospital, University Hospital Centre (Centre Hospitalier Universitaire [CHU]), Montreal, Quebec, Canada
[5]Academic Affairs, Holland Bloorview Children's Rehabilitation Hospital, Toronto, Ontario, Canada

Correspondence: Dr Faizal Haji, SickKids Learning Institute, Room 620, 525 University Avenue, Toronto, Ontario M5G 2L3, Canada. Tel: 00 1 647 972 8086; E-mail: faizal.a.haji@gmail.com

INTRODUCTION

For nearly 40 years, programme evaluation in the health professions has been shaped by the widespread adoption of the Kirkpatrick hierarchy.[1,2] Originally proposed as a framework for evaluation in human-resources training[3] presently many studies evaluating health sciences programmes use a variation of its 4 outcome-levels (participation; attitudes, knowledge and skills; behavior change; and systems-level impacts, such as improved patient outcomes).[4–6] The reason for the model's predominance is evident: it provides a clear taxonomy for making evaluative judgements, utilising an uncomplicated structure that places the outcomes of greatest interest at the top of the hierarchy.

In recent years, however, educational programmes in the health professions have come under increasing scrutiny. Much of the current consternation relates to the limited impact curricular interventions have had on primary outcomes of interest.[7] For instance, meta-analyses of continuing medical education programmes repeatedly show small effects on higher-level outcomes, such as physician behaviours and patient care.[8,9] The realisation that our programmes have demonstrated minimal effects on our intended outcomes leads to one of two conclusions: either little of what we do makes a difference, or our existing evaluation models are inadequate to capture the effects we are interested in.

However, the plethora of studies demonstrating 'no significant difference' may stem from a larger problem: namely, that the *questions* we are asking are inadequate. In the wake of the best evidence medical education (BEME) movement that has emerged in recent years,[2] we have defined the purpose of programme evaluation to be to place value on an activity,[10] or to demonstrate its 'merit or worth'.[11] Our ability to make such judgements rests principally on the evaluation of the effectiveness of our programmes, in which what we define as 'effective' is 'inescapably linked to the outcomes of our educational interventions'.[6] In other words, we consider a successful educational programme as one that has achieved its predetermined outcome(s).

The inevitable result is the adoption of outcomes-driven models of evaluation, such as the Kirkpatrick hierarchy, that are capable of generating evidence that prove our interventions 'work'.[12] However, an exclusive reliance on outcomes-based approaches as the *sine qua non* of evaluation is too narrow in scope and cannot account for the complexities of the health professions education context. It is not surprising, therefore, that in recent years an increasing number of published reports in the health professions education literature have utilised alternative models of evaluation that consider factors such as a programme's context, process and theory. In an attempt to generate a deeper understanding, these models embrace a reconceptualised view of the purpose of the evaluative act: it is not just about judging merit or worth, but also about generating reliable, valid and useful information for curriculum developers seeking to adapt programmes in the light of evolving contexts, and health professions education researchers seeking to generate knowledge that can 'inform the efforts of others'.[13] This change in focus is in keeping with recent 'calls to action' published by prominent medical education scholars, who have articulated that we must move beyond the 'imperative of proof'[12] to focus on 'clarification studies'[14] that additionally ask how and why our interventions do (or do not) work, and seek to establish what else is happening when our programmes are implemented.

As a result, health professions education has found itself at a turning point: although new trends in programme evaluation are emerging, they have not yet made their way into the mainstream approach, which is still dominated by outcome-based evaluation. Interestingly, a similar paradigm shift occurred within the discipline of educational programme evaluation nearly 50 years ago. In this paper, it is our intent to support the evolution of programme evaluation currently underway in the health professions, by drawing on and learning from the evolution of programme evaluation within this broader field. To do so, we begin by tracing the historical roots of the discipline of programme evaluation, demonstrating parallels with the dominant, outcomes-based approach to evaluation in the health professions. Next, we review the evolution of contemporary paradigms within this field, both to highlight key lessons learned by evaluation theorists and to draw attention to emerging scholarship in the health professions that has used these paradigms, in order to provide examples that readers may draw upon to inform the evaluation of their own programmes. Lastly, we will present the essential elements that we believe future evaluative efforts should incorporate, in order to address the questions 'how', 'why', and 'what else happened', allowing us to move beyond 'did it work?'

HISTORICAL ROOTS OF EDUCATIONAL PROGRAMME EVALUATION: A PARALLEL TO THE HEALTH PROFESSIONS

Although early evaluation efforts date back to the late 1800s, programme evaluation as a discipline developed earliest and most intensely within the field of education. Evaluation scholars often cite Ralph Tyler's coining of the term 'educational evaluation' in the 1930s and 1940s as a landmark event in the development of the modern profession and discipline.[15,16] Following his work on the Eight-Year Study, Tyler came to view evaluation as the appraisal of an educational programme's quality.[17] Defining quality in terms of a programme's effectiveness in achieving its predetermined goals,[18] the Tylerian paradigm called for a linear, hierarchical approach that compared planned outcomes with objectives defined *a priori*.

The Russian launch of the Sputnik satellite in 1957 precipitated a national crisis in the USA. Reflecting an effort to compete on a global scale, the National Defense Education Act was passed, leading to the rapid expansion of educational programmes in math, science and foreign languages.[19] The Tylerian approach was adopted to define objectives for this new curriculum and national standardised tests were created to better reflect these objectives and curricular content. With the infusion of capital into education came a desire to evaluate the effectiveness of these programmes.[19] Influenced by the writings of Campbell and Stanley,[20] large-scale field experiments were used to evaluate the newly developed curricula with respect to planned outcomes.

Despite best efforts, these hugely expensive and widely attempted experimental studies often demonstrated 'no significant difference'.[15] Even when significant results were reported, little information was provided on the nature of the programme and the manner in which it was implemented. Analogous to the results of 'grand curricular experiments'[21] in medical education that have recently been called into question, by the late 1960s it was apparent to leading evaluation scholars that this approach neither provided insight to decision makers on how to improve programmes, nor adequately addressed questions about a programme's 'effectiveness'.[19]

The reader may now be experiencing an uneasy sense of familiarity. This is not your imagination: programme evaluation in the health professions has followed a trajectory parallel to that witnessed by our counterparts in educational program evaluation, albeit 50 years later. Focusing evaluations on demonstrating the 'effectiveness' of our interventions through an outcome-oriented approach has caused us to fall into the same pattern as evaluation scholars of the past; as a result, we too have failed to generate meaningful understanding of the factors that lead to the success or failure of a programme and the interactions of these factors within the complex, multivariate system that epitomises health professions education. Fortunately, many medical educators have recognised this alarming trend, and novel approaches to evaluation are emerging within health professions education scholarship. In the next section, we will trace the evolution of some contemporary paradigms in programme evaluation that have emerged over the last 40 years, highlighting both the lessons that can be learned from these approaches, as well as examples of how they have been used in the evaluation of educational programmes in the health professions.

EVOLUTION OF CONTEMPORARY PARADIGMS IN PROGRAMME EVALUATION: UNDERSTANDING HOW AND WHY PROGRAMMES WORK

As a result of the shortcomings of traditional evaluation approaches that developed in the Tylerian age, many evaluation scholars became disheartened with the status quo. Consequently, beginning in the late 1960s a number of paradigm shifts occurred in the theory and practice of educational evaluation. It is not our intent to provide an exhaustive overview of the many programme evaluation theories and models that emerged during this critical period of development. Rather, as we attempt to catch up on 50 years of evolution, we hope to selectively highlight those approaches that illuminate both the essential elements that must be included in our evaluative efforts, as well as the shortcomings of individual evaluation models that prompted the development of subsequent paradigms. For those interested, we would recommend for further reading the eloquent and detailed typology of programme evaluation provided by Alkin and Christie,[16] and the comprehensive review of programme evaluation theory, models and applications by Stufflebeam and Shinkfield.[15]

One of the first paradigm 'shifts' to occur during this critical period was catalysed by a group of evaluation theorists who believed that the primary purpose of evaluation was to provide improvement-oriented, user-centred information to stakeholders for the purposes of decision making. A number of evaluation models, including Daniel Stufflebeam's CIPP (*c*ontext, *i*nput, *p*rocess, *p*roduct) model,[22] Robert Stake's responsive evaluation,[23,24] and Michael Patton's utilisation-focused evaluation (UFE),[25] developed within this paradigm. For these evaluators, the educational context in which a programme operates plays a significant role in the articulation of evaluation questions, evaluation methods and interpretation of evaluation findings. In addition, by focusing evaluation questions on the needs of programme stakeholders, these models bring to the fore the importance of considering the educational processes involved in programme implementation, in addition to measuring outcomes.[24] For instance, in Stufflebeam's CIPP model,[22] the evaluation of educational processes (the first 'P') involves asking 'are we doing it correctly?' (or, to put it another way, 'Did we do what we said we would?') to determine whether programmes are delivered in the manner in which their designers intended. A number of evaluations using these models, or similar approaches, have appeared in the health professions education literature in recent years.[24,26–29] For example, a modification of Robert Stake's responsive evaluation model was used by Curran *et al.*[24] to evaluate a clinical skills and assessment programme for doctors working in rural communities in eastern Canada. The ensuing meta-evaluation demonstrated the utility of this approach for gathering opinions from an array of programme constituents, which allowed for the articulation of conflicting needs, opinions and values in the judgement of programme outcomes. Additionally, Steinert *et al.*[26] used the CIPP model to evaluate a faculty development programme for teaching and assessing professionalism. Their evaluation of programme context and process, which consisted of surveys and informal interviews with stakeholders, revealed that the identification of core concepts, the provision of a structured framework for teaching and evaluating professionalism, and the analysis of case vignettes in small groups were particularly useful for participants.[26] It would appear that these factors contributed to the positive outcomes observed by the authors, which included an increase in medical education activities for trainees directed at professionalism, as well as the incorporation of these concepts into the clinical practice and teaching of faculty participants.[26] Utilising a similar model known as the 3P (*p*resage, *p*rocess, *p*roduct) model, Reeves and Freeth[27] demonstrated the vital importance of contextual factors in relation to a lack of continuity of leadership and engagement of senior management in the failed long-term viability of an in-service interprofessional education programme for community mental health teams. In this way, such context- and process-oriented approaches can provide insight into *how* programmes operate to bring about (or fail to bring about) their intended outcomes.

One of the limitations of focusing evaluation solely on the decision-making needs of stakeholders is that although such evaluations may provide valuable information regarding programme processes, they fail to inform us about *why* programmes work. Thus, if programme processes fail to produce their intended effects, there is little explanation for this result. Fortunately, this issue is addressed by the theory-based evaluation paradigm, which emerged during the mid-1970s and 1980s through the work of Huey-Tsyh Chen, Stewart Donaldson and others.[30] The main purpose of theory-based evaluation is to understand why a programme is succeeding or failing, for the purposes of programme improvement and 'knowledge construction'.[31] In other words, theory-based evaluations seek to unpack the 'black box' by identifying mechanisms that mediate between programme processes and intended outcomes[30] in the hope that these findings can be generalised to similar programmes in similar situations.

To illustrate this type of evaluation, consider the law of universal gravitation. In Newton's classic observation of an apple falling from a tree, the final position of the apple can be viewed as an 'outcome'; the characteristics of the apple, the tree and the earth as the 'context', and the path of the apple as it falls through the air as the 'process'. However, understanding each of these facets does not adequately inform us about *why* the apple falls. Only by articulating the existence of an attractive force (gravity) and its action upon the apple can we adequately explain the mechanism by which the apple gets from the tree to the ground. Similarly, proponents of theory-based evaluation argue that to generate an understanding of why programmes operate in the way they do, evaluators must articulate (and subsequently evaluate) a 'plausible and defensible'[32] conceptual framework (i.e. theory) that explains the mechanism by which programme processes lead to outcomes.

Thus, a theory-based evaluation would begin with an articulation of the conceptual framework that underlies programme design, based on programme designers' and evaluators' understandings of why the programme will work. In this way, the theory serves as a prediction and answers the question of what will happen. Based on this prediction, pertinent evaluation questions, variables and designs are identified and an implementation plan is generated. The evaluative phase first considers whether programme implementation is consistent with the articulated theory because when it is not, a 'failure of implementation' can occur.[30] Subsequently, through the simultaneous use of randomised experiments and advanced statistical techniques (such as structural equation modelling),[16] causal linkages between programme theory and intended outcomes are evaluated. Thus, when the implementation plan is carried out correctly and the desired results are achieved, the evaluator can conclude that the stated theory accounts for the mechanism linking the programme processes to the observed outcomes.

In this paradigm, the selection of an appropriate theory for a given programme is largely dependent on what the programme hopes to achieve. For instance, consider a programme that seeks to teach new nursing staff in the intensive care unit about appropriate care of central venous catheters with the aim of reducing line sepsis rates. In this example, the programme's objectives are to facilitate the learning of new staff and to promote behaviour change to ensure best practices are followed. Thus, theories concerning both learning and behaviour change are likely to be mechanistically relevant to this programme. As Hodges and Kuper argue,[33] three classifications of programme theory are applicable in the health professions context: bioscience theories (e.g. theories related to motor learning and cognitive load), learning theories (e.g. situated learning, adult learning theories and socio-cognitive theory), and socio-cultural theories (e.g. critical and politico-economic theories). The understanding garnered from evaluations that are built on these theories can further our understanding of how students learn and what motivates behavioural change within complex systems. Thus, the applicability of this approach to health professions education is clear and, not surprisingly, many scholars have already called for an increase in theory-based approaches.[1,34,35]

Interestingly, in an attempt to operationalise Chen's model of programme theory[32] to the evaluation of a medical education fellowship programme, Parker *et al.*[34] demonstrated that the consideration of programme theory may unearth previously undefined tensions within programme elements. In a creative approach to address this tension, the authors engaged a secondary process known as 'polarity management',[34] drawn from the organisational development literature, to deepen their understanding of this tension and how it related to the core strengths of the programme. In doing so, they were able to unravel another layer of the programme's theory and thus a better understanding of how not only intended, but also unintended and emergent outcomes of the programme came to be. In the next section, we demonstrate that the consideration of emergence is vital to gaining a deep understanding of the complex programmes that epitomise health professions education, and how emergent outcomes, processes and theories have been addressed in programme evaluation to date.

EMBRACING COMPLEXITY: EVALUATING EMERGENCE TO ESTABLISH WHAT (ELSE) HAPPENED

The paradigms we have discussed thus far rely heavily on articulating programme theory, processes and outcomes in advance of implementation. Based on the assumption that events occurring under the same conditions will produce similar results, this approach attempts to generate simplified, generalisable 'truths' that can be broadly applied to curriculum-level interventions.[12] Yet educational interventions are not singular entities;[36] they consist of a myriad of dynamic components that interact in complex, non-linear ways, influenced by ever-changing contexts, in which unpredictability is the rule. Thus, the lack of consideration of what is *actually* happening in the moment (as opposed to what we *predict* will happen based on our goals) limits our capacity to account for complexity. In other words, as these paradigms fail to ask 'what (else) happened?', they cannot capture the unintended processes and outcomes that emerge as programmes are operationalised. We need only consider the impacts of the hidden curriculum in undergraduate medical education to understand how a failure to capture unintended effects can lead to evaluations that may well miss the mark.

Fortunately, the need to capture 'emergence' was recognised by evaluation theorists early in the evolution we have been tracing. Among the first to do so was Michael Scriven, who presented his formulation of 'goal-free evaluation' in the early 1970s.[16] Scriven's perspective is that the primary purpose of evaluation is to place value on a programme, or to judge its merit or worth.[37] However, he argues that to adequately do so, evaluators must consider the *actual* effects of a programme, whether they were intended or not.[37] Scriven goes so far as to suggest that, analogous to blinding in clinical trials, all evaluations should be completed by external experts who have no knowledge of the goals of a programme, so that they may render an impartial evaluation based on observed outcomes.[37] Although this is an extreme view (and

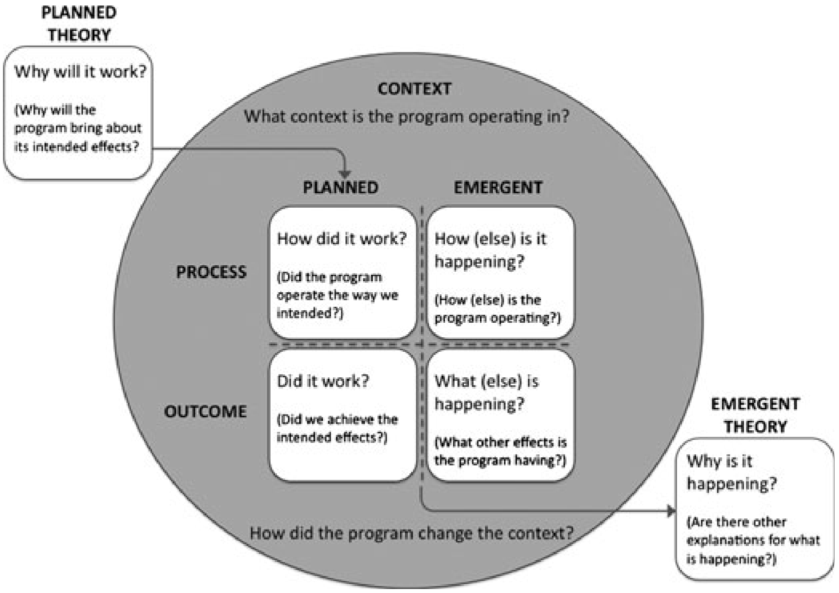
PLANNED THEORY
Why will it work?
(Why will the program bring about its intended effects?)
CONTEXT
What context is the program operating in?
PLANNED
EMERGENT
PROCESS
How did it work?
(Did the program operate the way we intended?)
How (else) is it happening?
(How (else) is the program operating?)
OUTCOME
Did it work?
(Did we achieve the intended effects?)
What (else) is happening?
(What other effects is the program having?)
How did the program change the context?
EMERGENT THEORY
Why is it happening?
(Are there other explanations for what is happening?)

Figure 1 Evaluation essentials. This figure delineates the seven essential elements of the programme evaluation process outlined in this paper. Alongside each element, we articulate the associated evaluation question. In our formulation, evaluation begins at the point of programme conception, with the articulation of a planned theory. This helps programme designers articulate how a programme should operate (planned processes) in order to achieve planned outcomes, which are evaluated at programme completion. To capture *what (else) happened*, evaluation during programme delivery must also consider emergent processes and outcomes *in articulo*. The dashed line denotes that these relationships are not linear; both planned and emergent processes can lead to planned or emergent outcomes. In addition, the circle represents the larger context in which the programme operates, which both influences and is modified by the programme. The evaluation culminates with a careful search for additional explanations (emergent theory) for why emergent processes and outcomes came to be, and to articulate alternative mechanisms (beyond planned theory) behind planned processes and outcomes

depending on the reader's epistemological position, a point of controversy with respect to whether 'impartial' evaluation is ever possible), it highlights the importance of considering not only planned outcomes, but unplanned (i.e. emergent) ones as well when making judgements regarding a programme's effectiveness. The importance of this concept should be readily apparent to health care professionals, as the analogous circumstance of judging the effectiveness of a proposed treatment without considering its potential side-effects would obviously be inadequate.

Emergence has also gained prominence in the evolution of the UFE paradigm. This stems principally from Michael Patton's work in the mid-1990s on the concept of 'developmental evaluation'.[38] As its name suggests, this approach evolved from a desire to better serve the needs of programme designers, who were interested not in making judgements about their programmes, but in understanding the implications of what they were developing. These designers 'never expect to arrive at a steady state of programming because they're constantly tinkering as participants, conditions, learnings, and contexts change…[and] no sooner do they articulate and clarify some aspect of the process than that very awareness becomes an intervention and acts to change what they do'.[38]

Recognising the limitations of traditional UFE in addressing the needs of programme designers, Patton articulates an approach in which there is no temporal distinction between the development and evaluation process. The evaluator becomes an integral part of the development team,[38] helping designers to monitor both processes and outcomes as they emerge from the evolving, rapidly changing environment; evaluation literally occurs *in the moment*. As the focus is on providing whatever information is needed for the development process, the choice of methodological approach is based on what is most appropriate for the situation. This can include quantitative methods (such as structured, numerically anchored stakeholder surveys), qualitative methods (e.g. interviews, focus groups and structured stakeholder conversations such as those defined by the ORID [*o*bjective, *r*eflective, *i*nterpretive, *d*ecisional] model[39]) or, in most cases, mixed methods which combine these approaches.

Although this approach brings up a multiplicity of new ideas, one of the most salient to our discussion is the notion of capturing adaptation. Patton's formulation[38] provides an avenue by which to answer not only the question of whether we did what we said we would, but also to establish how (else) the programme is operating. That is, how did participants, programme designers, other stakeholders and the *context itself* adapt to the programme, and what additional processes emerged as a result? For an illustration of this, reconsider the example of the programme to teach new nursing staff how to care appropriately for central venous catheters. If a number of experienced staff were to express interest in completing the programme as a 'refresher', programme designers adapting to this change in context might need to vary the delivery of the programme to accommodate a larger number of participants. It would be important to capture these 'emergent processes' in the evaluation of the programme so that future programmes could be planned to include both new and experienced staff.

The evolution of theory-based evaluation has also embraced the notion of emergence in recent years. In the late 1990s, Ray Pawson and colleagues articulated an approach known as 'realist evaluation', which acknowledges and accommodates the 'messiness of real-world interventions'.[36] In line with the notion that one can never step twice into the same river (i.e. the context in which our interventions operate is changed merely by our interventions operating within them), the realist view contends that relying solely on programme theories articulated *a priori* is inadequate to explain the emergent processes and outcomes that result from programme implementation. From our example of the universal law of gravity, consider the instance of a leaf falling from the tree, rather than an apple. It is conceivable that instead of falling directly to the ground, a leaf might land 20 feet away (an emergent outcome). Upon observing the leaf, we recognise that although it eventually falls to the ground, it takes a tortuous route, floating through the air before reaching its final resting place (an emergent process). Our planned theory (gravity) cannot explain how the leaf came to rest so far away from the tree.

As such, relying solely on planned theory is inadequate because it leaves the evaluator with no recourse when planned outcomes are not achieved (in the absence of an 'implementation failure' to explain this finding). Even when intended outcomes are observed, there are often multiple mechanisms that can explain how this came to be. In realist evaluation, articulation of a programme theory is thus viewed as an 'iterative, explanation-building process'.[36] As the programme is implemented, the evaluator observes processes and outcomes as they occur, identifying and constructing one or more theories that might explain the mechanism between them. This formulation is constantly compared against emerging data and iteratively revised to ensure that it explains the findings. As the programme theory is articulated based on what is happening *in the moment*, we have termed this 'emergent theory'. However, it is important to note that a realist evaluation that *only* considers emergent theory may not be the most efficient approach. When programme designers and evaluators have a sense of (some of) the mechanisms at work within a programme, the *combined* use of planned and emergent theory may provide a better understanding. Returning to our example of the leaf falling from the tree, using this approach we would be able to consider the possibility of other forces, such as wind or air resistance (our emergent theory) acting in concert with gravity (our planned theory) on the leaf to better account for 'what (else) happened'.

Although research evaluating 'emergence' is uncommon in health professions education, some recent studies have attempted to apply these concepts. For instance, protocols and reports utilising realist evaluation are beginning to appear in the health professions literature.[40,41] In a related approach, Parker *et al.*,[1] developed a strategy for evaluating a clinician-scientist training programme using both the logic model and grounded theory methodology to capture planned processes, as well as planned and emergent outcomes. Using the evaluation data generated and by conducting a thorough literature search, the authors were also able to articulate an emergent programme theory that incorporated the emergent outcome (identity change for participants in the programme), along with the predetermined processes and outcomes outlined in the logic model.[1]

LESSONS LEARNED: A RECONCEPTUALISATION OF PROGRAMME EVALUATION IN THE HEALTH PROFESSIONS

As the reader will now realise, we in the health professions have much to learn from the evolution that has occurred in the field of programme evaluation in the last few decades. It is evident that an evaluation strategy that focuses solely on planned outcomes will not be sufficient to meet the needs of programme evaluators in the health professions moving forward; thus, we must continue to incorporate contemporary evaluation paradigms into our evaluation efforts because failure to do so will cause us to promote a linear, inflexible approach to evaluation[42] that will not provide the necessary information for decision making, knowledge generation, or judgement of the merit or worth of our interventions. However, it is not the intent of this paper to set aside the evaluation of planned outcomes and the question of whether an intervention 'worked'. Instead, we hope to promote critical reflection on where this question should sit in the evaluative process, and ensure that we are gathering information that can help us to answer the other, equally important, questions of how something worked, why it did and what (else) happened.

In support of this, we have traced the evolution of programme evaluation as a discipline over the last 50 years. To prevent us from reliving this history, we have sought to highlight the lessons that have emerged from this evolution. This has enabled us to draw links between the questions we are asking and the essential elements of a programme that must be evaluated to allow us to answer them. We have summarised these elements, and the relationships among them, in Fig. 1. In brief:

1 To address the question of *how* a programme works, evaluators must capture not only programme *outcomes*, but also programme *processes*.
2 An understanding of *why* a programme works requires an evaluation of programme *theory* to elucidate the mechanisms that can explain how programme processes lead to programme outcomes.
3 To truly understand *what (else) happened*, it is not sufficient to rely solely on the articulation of planned theory, processes and outcomes. The evaluator must also capture *emergence*, in terms of both processes and outcomes, and generate emergent theory to explain what is occurring *in articulo* (in the moment).
4 It is also important to acknowledge that all programmes operate within an educational context, which must be considered for any evaluation to be complete. Furthermore, simply by virtue of programme delivery, this context will change; capturing this change is essential for the planning of future programme iterations.

It is important to recognise that although they are presented sequentially, the links between planned and emergent theory, processes and outcomes are far from linear. A planned process may lead to a planned outcome, but can just as easily result in an emergent (unintended) outcome. Similarly, an emergent process may result in an emergent outcome, or by some previously undefined mechanism, the planned outcome as well. This underscores the importance of both articulating a planned theory, which allows for evaluation to be designed to test the causal relationships between planned processes and outcomes, and seeking out emergent theory to explain what is occurring as the programme unfolds.

The reader should also note that it is not our intent to provide a prescriptive model or methodology of how to conduct a programme evaluation. Rather, we hope to initiate a conversation about what programme evaluation in the health professions *should* be. Analogous to the unpacking of a Russian nesting doll, we believe the programme evaluation process begins with an understanding of the transcendent concepts of *what* should be evaluated (i.e. the elements provided in Fig. 1), before the choice of which 'model' or 'method' to use can be made (the 'how should we do it?'). The conceptual 'layer' provides a lens through which to make our selection of the most appropriate evaluation strategy (i.e. the appropriate evaluation model). In turn, the model 'layer' focuses our attention on evaluation methods that will address these needs, be they qualitative, quantitative or mixed.

In truth, no single paradigm satisfies all of the requirements we have set forth: in each, some of the elements we have proposed are prioritised over others. Thus, the choice of an approach to programme evaluation is not so much a treasure hunt for the 'perfect model' as it is a reflective exercise in which the evaluator recognises the inherent biases associated with his or her selection and decides on the most appropriate *combination* of available approaches. In this way, the choice between models (and methods) that emerge from these paradigms should not be viewed as an 'either/or' choice, but rather as a 'both/and' selection. In fact, we are not alone in our desire to blur the lines; as the profession has evolved, many programme evaluators have embraced the characteristics of seemingly opposing views in an effort to meet their needs.[1] By adopting this philosophy, we can be judicious in supplementing our chosen approach with models from other paradigms in order to ensure that we will capture *all* of the relevant elements in our evaluations.

On a final note, although we have purposely avoided being prescriptive in this discussion, we would be remiss if we did not point out the few fundamental changes we feel need to be made to the way we conduct evaluations in the health professions. Firstly, instead of treating evaluation as a snapshot endeavour that occurs after programme delivery, we must consider programme evaluation as a process in and of itself. In parallel with the notion of 'programmes of assessment', which has appeared in the health professions literature of late,[43] we must similarly move towards 'programmes of evaluation'. In recognition of the maxim that the whole is greater than the sum of its parts, evaluations must involve multiple stakeholders, use multiple methods, and occur *throughout* the life of the programme (right from conception, through to planning, delivery and revision) to generate a holistic understanding not only of the programme and its effects, but also of how the context in which the programme operates is changed by its presence. To do so effectively, we must relinquish our bonds to the arbitrary distinction between the process of *development* and that of *evaluation*; instead, we must see these activities as two sides of the same coin, each of which informs the other in a continuous, iterative process that leads to incremental programme change. It is only through this incremental change process that we will generate a sound understanding of the relationships among the health professions context, our interventions and the outcomes we seek.

CONCLUSIONS

It is clear that programme evaluations using traditional 'outcomes-based' models are inadequate for the health professions context. Consequently, the scholarship in health professions education has begun to incorporate alternative evaluative approaches capable of addressing the inherent complexity of the field. This is not unlike changes that occurred in the broader discipline of programme evaluation over 50 years ago, and as demonstrated in this paper, we have much to learn from the evolution of contemporary paradigms within this field. From this evolution, we have been able to highlight the importance of evaluating educational context, planned and emergent processes and outcomes, and planned and emergent theory. These elements have allowed us to address not only the fundamental question of *whether* our programme worked, but also the issues of *how* it worked, *why* it worked and *what (else) happened*. Only by embracing all of these elements can we hope to begin to understand the relationships among our interventions and the outcomes we are seeking. In time, and with this understanding, we may be able to build interventions that have the potential to achieve our ultimate goal: improved patient care.

Contributors: FH and M-PM worked collaboratively to develop the primary content of this paper. The initial draft of this manuscript was written by FH and revised for important intellectual content and clarity by M-PM and KP. KP developed a preliminary version of the framework presented in the final section, which was elaborated and revised by FH and M-PM. All authors approved the final manuscript for publication.
Acknowledgements: we would like to thank Jocelyn Lockyer, Department of Community Health Sciences, University of Calgary, Calgery, AB, Canada; Glenn Regehr, Centre for Health Education Scholarship, University of British Columbia, Vancouver, BC, Canada; and Rabia Khan, School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada, for their constructive comments, which were invaluable in improving this manuscript.
Funding: at the time of writing, FH was supported by the Royal College of Physicians and Surgeons of Canada Fellowship for Studies in Medical Education and M-PM was supported by a Joint Fund Award from CHU Sainte-Justine.
Conflicts of interest: none.
Ethical approval: not applicable.

REFERENCES

1 Parker K, Burrows G, Nash H, Rosenblum ND. Going beyond Kirkpatrick in evaluating a clinician scientist programme: it's not 'if it works' but 'how it works'. *Acad Med* 2011;**86** (11):1389–96.
2 Yardley S, Dornan T. Kirkpatrick's levels and education 'evidence'. *Med Educ* 2011;**46** (1):97–106.
3 Kirkpatrick D. Evaluation of training. In: Craig RL, Bittel LR, eds. *Training and Development Handbook*. New York, NY: McGraw-Hill 1967;87–112.
4 Dixon J. Evaluation criteria in studies of continuing education in the health professions: a critical review and a suggested strategy. *Eval Health Prof* 1978;**1** (2):47–65.
5 Barr H, Freeth D, Hammick M, Koppel I, Reeves S. *Evaluations of Interprofessional Education: A United Kingdom Review of Health and Social Care*. London: CAIPE/BERA 2000.
6 Belfield C, Hywel T, Bullock A, Eynon R, Wall D. Measuring effectiveness for best evidence medical education: a discussion. *Med Teach* 2001;**23** (2):164–70.
7 Colliver JA. Effectiveness of problem-based learning curricula: research and theory. *Acad Med* 2000;**75** (3):259.
8 Marinopoulos SS, Dorman T, Ratanawongsa N, *et al*. *Effectiveness of Continuing Medical Education*. Evidence

Report/Technology Assessment No. 149. Rockville, MD: Agency for Health Research and Quality 2007;1–69.

9 Forsetlund L, Bjørndal A, Rashidian A, Jamtvedt G, O'Brien MA, Wolf F, Davis D, Odgaard-Jensen, Oxman AD Continuing education meetings and workshops: effects on professional practice and health care outcomes. *Cochrane Database Syst Rev* 2009;**2**:CD003030.

10 Wilkes M, Bligh J. Evaluating educational interventions. *BMJ* 1999;**318** (7193):1269–72.

11 Cook DA. Twelve tips for evaluating educational programmes. *Med Teach* 2010;**32** (4):296–301.

12 Regehr G. It's NOT rocket science: rethinking our metaphors for research in health professions education. *Med Educ* 2010;**44** (1):31–9.

13 Eva KW. The value of paradoxical tensions in medical education research. *Med Educ* 2010;**44** (1):3–4.

14 Cook DA, Bordage G, Schmidt HG. Description, justification and clarification: a framework for classifying the purposes of research in medical education. *Med Educ* 2008;**42** (2):128–33.

15 Stufflebeam DL, Shinkfield AJ. *Evaluation Theory, Models, and Applications*, 1st edn. San Francisco, CA: Jossey-Bass 2007;32–41.

16 Alkin M, Christie C. An evaluation theory tree. In: Alkin M, ed. *Evaluation Roots: Tracing Theorists' Views and Influences*. Thousand Oaks, CA: Sage Publications 2004;12–65.

17 Goldie J. AMEE education guide no. 29: evaluating educational programmes. *Med Teach* 2006;**28** (3): 210–24.

18 Tyler R. General statement on evaluation. *J Educ Res* 1942;**35**:492–501.

19 Madaus G, Stufflebeam D. Programme evaluation: a historical overview. In: Stufflebeam DL, Madaus GF, Kellaghan T, eds. *Evaluation Models – Viewpoints on Educational and Human Services Evaluation*. Norwell, MA: Kluwer Academic Publishers 2000;3–18.

20 Campbell DT, Stanley J. *Experimental and Quasi-Experimental Designs for Research*, 1st edn. London: Wadsworth Publishing 1963;13–63.

21 Norman G. RCT = results confounded and trivial: the perils of grand educational experiments. *Med Educ* 2003;**37** (7):582–4.

22 Stufflebeam D. The CIPP model for programme evaluation. In: Madaus G, Scriven M, Stufflebeam D, eds. *Evaluation Models: Viewpoints on Educational and Human Services Evaluation*. Norwell, MA: Kluwer Academic Publishers, 2000;279–318.

23 Stake RE, Fund JDR. *Evaluating the Arts in Education: A Responsive Approach*. London: Merrill Publishing 1975;33–58.

24 Curran V, Christopher J, Lemire F, Collins A, Barrett B. Application of a responsive evaluation approach in medical education. *Med Educ* 2003;**37** (3):256–66.

25 Patton MQ. *Utilization-Focused Evaluation*, 4th edn. London: Sage Publications 2008;35–98.

26 Steinert Y, Cruess S, Cruess R, Snell L. Faculty development for teaching and evaluating professionalism: from programme design to curriculum change. *Med Educ* 2005;**39** (2):127–36.

27 Reeves S, Freeth D. Re-examining the evaluation of interprofessional education for community mental health teams with a different lens: understanding presage, process and product factors. *J Psychiatr Ment Health Nurs* 2006;**13** (6):765–70.

28 Al-Khathami A. Evaluation of Saudi family medicine training programme: the application of CIPP evaluation format. *Med Teach* 2012;**34** (Suppl 1):81–9.

29 Singh MD. Evaluation framework for nursing education programmes: application of the CIPP model. *Int J Nurs Educ Scholarsh* 2004;**1**:13.

30 Weiss C. Theory-based evaluation: past, present, and future. *New Dir Eval* 1997;**76**:41–55.

31 Shadish WR, Cook TD, Leviton LC. *Foundations of Program Evaluation: Theories of Practice*, 1st edn. Newbury Park, CA: Sage Publications 1990;171–377.

32 Chen HT, Rossi PH. Evaluating with sense: the theory-driven approach. *Eval Rev* 1983;**7** (3):283–302.

33 Hodges BD, Kuper A. Theory and practice in the design and conduct of graduate medical education. *Acad Med* 2012;**87** (1):25–33.

34 Parker K, Shaver J, Hodges B. Intersections of creativity in the evaluation of the Wilson Centre Fellowship Programme. *Med Educ* 2010;**44** (11):1095–104.

35 Norman GR, Schmidt HG. Effectiveness of problem-based learning curricula: theory, practice and paper darts. *Med Educ* 2000;**34** (9):721–8.

36 Wong G, Greenhalgh T, Westhorp G, Pawson R. Realist methods in medical education research: what are they and what can they contribute? *Med Educ* 2011;**46**(1):89–96.

37 Scriven M. Prose and cons about goal-free evaluation. *Am J Eval* 1991;**12** (1):55–62.

38 Patton M. Developmental evaluation. *Am J Eval* 1994;**15**:311–9.

39 Stanfield R. *The Art of Focused Conversation*. Gabriola Island, BC: New Society Publishers 2000;17–29.

40 Ogrinc G, Batalden P. Realist evaluation as a framework for the assessment of teaching about the improvement of care. *J Nurs Educ* 2009;**48** (12):661–7.

41 Dalkin SM, Jones D, Lhussier M, Cunningham B. Understanding integrated care pathways in palliative care using realist evaluation: a mixed methods study protocol. *BMJ Open* 2012;**2** (4):1533.

42 Worthen B, Sanders JR. *Educational Evaluation: Alternative Approaches and Practical Guidelines*. White Plains, NY: Longman Publications 1987;73.

43 Dijkstra J, van der Vleuten CPM, Schuwirth LWT. A new framework for designing programmes of assessment. *Adv Health Sci Educ Theory Pract* 2010;**15** (3):379–93.

Received 12 June 2012; editorial comments to authors 26 July 2012; accepted for publication 21 September 2012