Evaluation Matters—He Take Tō Te Aromatawai 6: 2020
© New Zealand Council for Educational Research 2020
https://doi.org/10.18296/em.0053
http://www.nzcer.org.nz/nzcerpress/evaluation-matters

Evaluation in dynamic times: Skateboard, pushbike, or quad bike?

Heather Nunns

The turbulent and fluid environment in which we find ourselves due to the COVID-19 pandemic requires evaluative responses that facilitate learning, adaptation, and timeliness. This article examines the last of these—the need for timely evaluative information. Such information requires evaluators and their clients making trade-offs between what is desirable and what is feasible in a constrained time frame. Applying a light-hearted analogy—skateboard, pushbike, quad bike—three different evaluative approaches and the trade-offs that each involves are described. The notion of *adequacy for purpose* is then examined in terms of two dimensions: the level of certainty of the evidence that is required by the client, and the level of confidence required by the client in the evaluative claim/conclusion. The article demonstrates the need for evaluators to ensure their clients and other users of evaluative information understand the level of confidence and certainty that can be placed in it.

"We will keep doing well if we keep learning and adapting and working collaboratively. We've had to tackle issues as they arise and that's been key to our success to date."

Dr Ashley Bloomfield, Director General of Health, COVID-19 media briefing, 1 July 2020

Introduction

Daily life has become more unpredictable and perplexing owing to the COVID-19 pandemic. At the time of writing, the public, private, and not-for-profit sectors in Aotearoa New Zealand are continuing to work out how to operate in this unique and fluid environment. The reflection by Dr Ashley Bloomfield about the country's COVID response to date describes some of the attributes required in this turbulent context—learning, adaptation, collaboration, responsiveness, and timeliness. If these are the attributes required for government agencies and other organisations in the pandemic world, then they are the priorities for our practice as evaluators. This article focuses on the last of these attributes—timeliness.

The demand for timely evaluation is not new. Some 10 years ago, I investigated five rapid evaluation and assessment methods (REAM) (Nunns, 2009) in response to the public sector's demand for quicker evaluation findings. These methods are designed for quick turnaround of findings required, for example, in humanitarian emergencies and other situations where just-in-time information is vital. More recently, developmental evaluation (Patton, 2011) has become the approach of choice for evaluators working with emergent interventions operating in dynamic contexts where ongoing evaluative feedback, learning, and adaptation are critical.

During a Zoom get-together in the lockdown with my KEA colleagues (KEA being the favoured name for a group of evaluators living on the Kāpiti Coast), I was reminded about the REAM paper. On re-reading it, I realised I had not adequately described the compromises or trade-offs that REAM methods involve. This article attempts to address this shortcoming by distinguishing between *skateboard, pushbike,* and *quad bike* approaches to evaluative activity. These terms are used in a light-hearted manner to illustrate differences in evaluation response—I have no intention of adding these terms to the already vast lexicon that is professional evaluation. Plus, in the interest of succinctness, I have chosen an arbitrary limit of three types of evaluative approach.

In this fast-moving COVID world, we aim to provide decision makers with the information they need when they need it. This involves compromises between what is desirable, and what is feasible and attainable given time and funding constraints. Applying our analogy, do we respond with a skateboard, or a pushbike, or a quad bike evaluative approach? To explain this transport analogy further—a valued colleague who reviewed a draft of this article pointed out that mountain and electric bikes can cost as much as a quad bike. In the interests of clarification, the pushbike referred to here is the basic variety. Further, the colleague expressed concern about the safety of quad bikes. Yes, a good point. In the interest of our clients' safety, the quad bike is fitted with a safety frame. Having clarified that our analogy is fit for purpose, let's move on.

Skateboards allow their riders to move fast and create agile movement. They are lightweight with few moving parts, making them an uncomplicated means of transport. The *skateboard evaluative approach* will be the more light-handed, nimble, and faster option of the three approaches. Unless a significant evaluator resource is available to do the work in a short timeframe, this response is likely to rely on secondary data, employ rapid data collection methods such as REAM, involve limited qualitative data, and use quick turnaround reporting approaches (Bamberger et al., 2006).

In contrast, a pushbike's frame (designed for strength and safety) supports mechanisms for steerage and braking to provide control and direction. While a bike may not move as fast or be as agile as a skateboard (professional bikers excepted), it has greater durability and therefore can be relied on for longer distance travel. The *pushbike evaluative approach* aims to balance rigour with speed. This invariably involves deciding between what is essential and what is desirable, given time constraints. Compromises are likely to be required, such as narrowing down the evaluative activity's focus, trading off depth with coverage, restricting the sample frame, and reducing the analytics.

The quad bike is a mainstay of rural Aotearoa. While a quad bike is an expensive purchase, the buyer can be reassured that they are getting a robust product for use in challenging conditions that meets legislated safety requirements. The *quad bike evaluative approach* is for the tougher terrain and/or the deep dive. This evaluation response is appropriate for the more political and/or complex evaluand, the dynamic context, the difficult to reach, and/or vulnerable programme participants (Smith, 1981). The high-stake nature of the tough terrain/deep dive requires an evaluative approach that provides trustworthy information. The quad bike evaluation design is likely to be grounded in a conceptual framework, employ multiple or mixed methods, involve baselines and/or comparison groups, have larger qualitative sample sizes and differentiated sampling frames, and employ more sophisticated analytical processes than the other two evaluative approaches.

Adequacy for purpose

When considering trade-offs between what is desirable and what is feasible in a constrained timeframe, how do we decide what is the most appropriate evaluative response—skateboard, pushbike, or quad bike? The answer is provided in accuracy standard A2 of the Program Evaluation Standards (Joint Committee on Standards for Educational Evaluation, 2011): "Evaluation information should serve the intended purposes and support valid interpretations." In other words, our proposed evaluative response should be adequate (valid) for the purpose for which the evaluative information will be used. The rest of this article attempts to unpack the notion of adequacy for purpose by discussing two dimensions: level of certainty required in the evidence, and level of confidence required in the evaluative claim/conclusion.

Dimension one: Level of certainty required in the evidence

Drawing on insights from the discipline of law, Smith (1981) distinguishes different levels of certainty of evidence relevant to health evaluations—suggestive evidence, preponderant evidence, and conclusive evidence. Taking a pragmatic perspective, Smith argues that the many different purposes for which evaluative information is used do not require the same level of certainty in the evidence collected. For example, ad hoc or just-in-time decision making may be based on suggestive evidence (described as the weakest form of evidence) which establishes that something is *plausible*.

Other purposes of evaluative information (e.g., lower level information-based decision making) may require more certainty. Preponderant evidence provides greater weight of evidence indicating the evaluative claim in question is *more possible*. However, while preponderant evidence is stronger, it is refutable.

Lastly, more consequential purposes of evaluative information (e.g., decision making about whether to continue a high value or contentious policy) requires conclusive evidence which Smith describes as *the beyond all reasonable doubt* standard of evidence (p. 274). Smith notes that evaluations of interventions involving children or vulnerable/at-risk adults require conclusive evidence. Similarly, evaluations of interventions with uncontrollable effects (e.g., the potential for unintended negative outcomes) require conclusive evidence to protect intervention recipients.

Smith questions whether the level of certainty of evidence needed by evaluation commissioners is explicitly addressed with them during scoping and design discussions. The following cautionary comment illustrates why this conversation is important: "The less certain the evaluation evidence is, the greater the probability of reaching an erroneous evaluative conclusion" (Smith, 1981, p. 277).

This statement highlights a fundamental point about how evidence is used to make evaluative claims. Specifically, the quality of evidence is directly correlated with the robustness of the evaluative claim/conclusion produced from it. Schwandt (2009, p. 201) identifies three properties of evidence that determine its value for making evaluative inferences, as follows. Schwandt emphasises that the assessment of each of the three properties is contextual and circumstantial. The assessment depends on factors including the perspectives of the users of the evidence, the kind of evaluative claim the evidence will be used to support, and the purposes for which the evidence will be used. The three properties are:

1. Relevance—does the evidence bear directly on the evaluative claim in question?
2. Credibility—can we believe the evidence?
3. Probative (inferential) force—how strongly does the evidence point towards the evaluative claim being considered?

Dimension two: Level of confidence required in the evaluative claim/conclusion

Evidence that is assessed by evaluation users as having these three properties is more likely to engender confidence in the trustworthiness (robustness) of the evaluative claim/conclusion for the purposes for which it will be used. This is now discussed further.

Examining REAM studies, Bamberger et al. (2006, p. 76) note that the speed with which such studies are undertaken means that "most of these rapid applications do not systematically address the increased threats to validity to which the findings may be subjected and there is a need … to assess the trade-offs between time, quality and validity".

The term *validity* describes "the soundness or trustworthiness of the inferences that are made from the results of the information gathering process" (Joint Committee on Standards for Educational Evaluation, 2011). I do not intend to discuss the types of quantitative and qualitative validity. For the purposes of this article, quad bike evaluative responses provide the most opportunity to strengthen validity through triangulation of the collected evidence, defined as: "a means of checking the integrity of the (evaluative) inferences one draws. It can involve the use of multiple data sources, multiple investigators, multiple theoretical perspectives, and/or multiple methods" (Schwandt 2007, p. 298).

Therefore, providing appropriate validity measures are applied, a quad bike evaluative approach will provide a stronger level of confidence in the evaluative claim/conclusion than a pushbike approach.

Trade-offs are an inevitable part of evaluation practice

Evaluators are pragmatic professionals. We know full well that evaluative findings that are too late will be consigned to the bottom drawer. We are used to having to compromise when working out how best to produce the required evaluative information within time and funding constraints. This article has highlighted the importance for evaluators to be explicit with evaluation commissioners and users about the implications of such trade-offs for the certainty and confidence that can be placed on evaluative information and the consequent limitations on how it can be used. This is particularly relevant for the environment in which we are now operating—we need to ensure that the demand for quick turnaround of evaluative findings does not undermine the credibility of evaluative information produced by Aotearoa evaluators. Smith (1981, p. 274) is unequivocal about our obligation: "evaluators have a professional responsibility to ensure that the recipients (of evaluative information) understand the level of confidence and certainty that can be placed in (it)".

Concluding remark

The demand for real-time evaluative information is likely to accelerate in the fast-changing environment triggered by COVID-19. Amidst the urgency and pressure involved in delivering timely information, we must be mindful of the need to deliver fit-for-purpose evaluative information, and to be explicit to evaluation commissioners and other users about how much confidence and certainty they can place in it.

References

Bamberger, M., Rugh, J., & Mabry, L. (2006). *Real world evaluation: Working under budget, time, data and political constraints*. Sage.

Joint Committee on Standards for Educational Evaluation. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Sage.

Nunns, H. (2009). Responding to the demand for quicker evaluation findings. *Social Policy Journal of New Zealand*, *34*, 89–99.

Patton, M. Q. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. The Guildford Press.

Schwandt, T. A. (2007). *The SAGE dictionary of qualitative inquiry* (3rd ed.). Sage.

Schwandt, T. A. (2009). Toward a practical theory of evidence for evaluation. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 197–211). Sage. https://doi.org/10.4135/9781412995634.d18

Smith, N. L. (1981). The certainty of judgments in health evaluations. *Evaluation and Program Planning*, *4*(3–4), 273–278. https://doi.org/10.1016/0149-7189(81)90028-8

The author

Heather Nunns (PhD) is the principal of Analytic Matters, a public-sector research and evaluation consultancy located on the Kāpiti Coast. Her work interests include connecting evaluation theory with the pragmatics of evaluation practice; and exploring how the practice of evaluative reasoning can be strengthened to support the reputation of Aotearoa evaluators as the “go to people” for insightful thinking.
Email: heather@analyticmatters.co.nz