Dealing with complexity in establishing outcomes and impact of innovation in education

Russell Tytler

Deakin University

tytler@deakin.edu.au

There is increasing pressure within education policy development to implement 'evidence-based' curricula and pedagogy. For funding of educational innovation this can amount to calls for 'gold standard' experimental designs, and for assessment a call for standardized testing. The danger however is that these methodological pressures can lead to concentration of policy and practice around 'low hanging fruit' of pedagogies and learnings that are readily measured. This presentation will draw on experience with educational innovation at the system level, and in the classroom, to point out problems with superimposing controlling methodologies over complex systems. It will explore distinctions between judgments of process and product, short and long-term outcomes, and the possibility of devising evaluation methods that are rigorous, yet cognizant of the complexity of educational systems, schools, and classrooms.

The Gold Standard Debate: Methodology wars in Education Research

For some years now there has been a call on the part of US authorities, echoed in writing also in Australia, for education research to move towards a 'gold standard' methodology involving experimental methods with randomized controlled trials. For a period in the US such experimental research was a requirement for National Science Foundation funding. The main thrust of the argument for such a gold standard has been a concern with the rigor of education research and a desire to ensure that innovation in education proceeds only on rigorous evidential grounds. Hempenstall (2006, p. 83) for instance argued in the Australian context:

> Teaching has suffered both as a profession in search of community respect and as a force for improving the social capital of Australia, because of its failure to adopt the results of empirical research as the major determinant of its practice. …. The generally low quality of much educational research in the past has made the process of evaluating the evidence difficult …

The methodological basis of the gold standard movement has been questioned on a number of grounds (Lawrenz & Huffman, 2006). First, it has been pointed out that science itself builds knowledge by far more varied methods than clinical trials, or experimental methods generally, including by observation and descriptive methods, modeling, and theory development drawing on a wide range of evidence. Second, there exists an enormous body of methodological writing establishing standards of rigor in a wider variety of research designs. For instance interpretivist studies and longitudinal designs (Tytler, 2009a) or case study methodologies all have their practices for establishing validity or trustworthiness. Third, an exclusive focus on experimental methods privileges one set of values only, around reductive nomothetic views involving the generation of general laws built on reproducible facts and outcomes, rather than a concern with understanding how individuals interpret and act in the world (Cohen, Manion, & Morrison, 2000). Further, limiting evaluation to "scientific" experimental approaches can privilege a regularity view of causation with a forced emphasis on the determination of patterns of relationships, rather than seriously considering how or why these patterns occur (Maxwell, 2004). Fourth, generating educational innovation builds on theoretical advances and a range of research designs to generate, validate and scale up new practices, with experimental designs relevant to only part of this process. Even with evaluation of innovation, there are strong arguments for mixed method designs rather than a reliance on experimental methods only (Lawrenz & Huffman, 2006). Finally, it must be acknowledged that at every level, from individual learning to teacher actions and classroom environments and beyond, education systems are complex, and difficult to reduce to categories simply and reliably described and measured in a manner assumed by gold standard advocacy.

Complexities in evaluating outcomes and impact

Nonetheless, there are strong expectations amongst policy makers and some researchers in education that more rigor needs to come into the field, with stronger reliance on experimental designs in evaluating innovation. In this paper I will describe cases of the evaluation of innovation and the challenge faced in providing evidence of improvement. It is not my intention to go over the ground of methodological justifications, which are well canvassed in the literature. Rather, I will use the cases to raise issues concerning a) complexities in conceptualizing and measuring outcomes and impacts of innovation, b) the appropriate comparison to be made in the evaluation, and c) how one might deal with complexity. Distinctions I will make, concerning the nature of outcomes, include:

- The grain size of outcomes
- Outcomes in the moment, and long term
- Process outcomes vs product outcomes
- Distinctions between inputs, outputs, outcomes and impact in evaluating innovation

The first distinctions, grain size and time scale, can be made by contrasting the outcomes of school education promoted in the Melbourne Declaration (Australian

Education Ministers, 2008), and those found in the Australian Curriculum: Science (ACARA, ND. The Melbourne Declaration include:

Successful learners:

- are able to think deeply and logically, and obtain and evaluate evidence in a disciplined way as the result of studying fundamental disciplines;
- are creative, innovative and resourceful, and are able to solve problems in ways that draw upon a range of learning areas and disciplines;
- are able to plan activities independently, collaborate, work in teams and communicate ideas.

Confident and creative individuals:

- have a sense of self-worth, self-awareness and personal identity that enables them to manage their emotional, mental, spiritual and physical wellbeing.

These are large grain size aims that speak to broad and deep individual capabilities and dispositions that are inevitably long term and "whole person" in nature. The ACARA outcomes by contrast include such understandings as:

- Mixtures, including solutions, contain a combination of pure substances that can be separated using a range of techniques;

or Inquiry skills such as:

- Collaboratively and individually plan and conduct a range of investigation types, including fieldwork and experiments, ensuring safety and ethical guidelines are followed.

These are shorter term, much smaller grain size outcomes, and much more readily measured using standard tests than the soft skills and higher order thinking implied by the Melbourne Declaration goals. The distinction between process and product outcomes is illustrated by the knowledge vs. inquiry skills ACARA outcomes. The planning outcome above seems to imply a capability to act within an investigation whereas the chemistry knowledge outcome is couched in "product" terms that can be presumably demonstrated by declarative means on a test. While there is nothing surprising in these distinctions between shorter and longer term, holistic capability and specific mastery, and process and product, it raises two questions about evaluation of educational outcomes. First, what would our assessment and evaluation traditions look like if we seriously focused on these longer-term, whole-of-person aims? Second, what does it tell us about the increasingly dominant focus on experimental methods demanding measures that are both reliable and reproducible, and which inevitably privilege these more prosaic, short-term outcomes? In this paper I use three cases to critically examine the nature of the difficulty of establishing specific educational outcomes in complex innovation environments, and to argue for a need to develop alternative ways of thinking about rigor in encompassing more substantive aims and outcomes in evaluating educational innovation.

Case 1: School Innovation in Science

The SIS research project developed a model of improvement in school science that focused on a set of *SIS Components of Effective Practice* and developed a set of processes by which the science teachers in a school, conceived of as a professional learning team, worked together to frame and develop changes in their pedagogy. Schools were supported by workshops focused on change processes, by network meetings, and by project officers and critical friends providing ongoing support (Tytler, 2007, 2009b).

The initial evaluation after the first year was specified in the contract, to be a random experimental design of SIS schools and a matched group, using a multiple choice test format focusing on science knowledge. The results for the first year did not show a significant difference in test scores, reflecting a number of complexities in operation of the innovation and the research design: a) the patterns of implementation of the SIS pedagogy were complex within the SIS schools, with some teachers embracing the change enthusiastically yet others resistant at least over that time frame; b) the control schools agreed to administer the tests on the proviso they were able to join the project in the second year, and there were a number of reports of teachers in these schools starting to implement the innovation informally in that first year, contaminating the research design; and c) there was a mismatch between the pedagogy based on expert practice, which emphasized inquiry and contextualized science curricula, and the multiple choice tests focusing on formal science understanding.

On the basis of that evaluation experience the team argued that the unit of measurement should be the classroom rather than the school. The conformity of the practice of individual teachers to the SIS Components were being mapped as part of the professional learning support processes, and the team was able to track changes in their Component Map scores and also compare test scores of classes where teachers showed high conformity (Hi SIS classes) to those who showed low conformity (Lo SIS classes). The result was a significant difference in test performance amounting to a full year of development, and a significant difference on an attitude survey. Over the 3 years of the project the mean SIS mapping score for all teachers rose substantially.

We argued that the demonstrated increase in component mapping scores together with the demonstration of better test outcomes of students in classes of teachers who scored high, was enough to demonstrate the success of the innovation. This method of comparison is of course not as tightly controlled as a gold standard experimental design but nevertheless offers some assurance of the value of the innovation. There was some significant bureaucratic opinion however that the project was not demonstrated to be successful. Indeed, in the third year for a variety of reasons the comparison fell below the significance level.

Nevertheless there were many and varied categories of evidence that the project was highly successful. These included evidence, through a survey of science coordinators, of a highly significant change in the way science staff worked together to establish a shared vision, to focus on teaching and learning rather than administrative processes, and to plan and implement curriculum with a focus on student learning outcomes. These processes are consistent with understandings in the

literature of effective professional learning processes. Further, the quality of student work, shown in teacher presentations at network meetings, was imaginative and of a high quality. There were many stories of the science staff in SIS schools being regarded as pedagogically innovative, such that principals reported using the model to energize other curriculum areas. Eventually the SIS Components were refined and transformed into a general framework that became the main plank of the state pedagogical direction for a number of years.

Thus, despite the difficulty of demonstrating rigorous evidence of improvement in student results, because of the complexity of the system, difficulties with practical and ethical issues of comparison, and the difficulty of tying down appropriate outcomes that could serve an experimental methodology, the project by its reputation was a catalyst for ongoing and significant effects. The question that arises is therefore: how might these other, multiple forms of evidence have been galvanized to tell the overwhelmingly positive story of the changes in school and teacher processes and in student learning, within a rigorous evaluation methodology? How could the wider impact of the project have been convincingly demonstrated?

Case 2: A representation construction inquiry pedagogy

This case concerns research over a number of funded projects, developing and validating a guided inquiry approach to teaching and learning science that involves students actively constructing and evaluating representations in response to teacher challenge. The research has been reported in many papers, and a book, and is based strongly in theoretical frames that link school science with practices in the scientific community. Evaluation of the outcomes of the project have included a variety of forms of evidence, both of process outcomes dealing with teacher and student demonstration of quality processes while engaged with the pedagogy, and product outcomes, identifying artifacts or demonstrated knowledge that flow from engaging with the pedagogy. Process outcomes include:

- Teacher perceptions and video records of high level class discussion associated with the approach
- Teacher testimony concerning improved student engagement with tasks
- Sophistication in student collaborative evaluation of quality of representations, revealed by video.
- Change in teachers' classroom strategies and epistemological beliefs

Product outcomes include:

- Student journal work and modeling activity of a quality beyond what would normally be expected
- Informal evidence through interviews, teacher perceptions and some test items, of student meta-representational sophistication beyond the norm
- Pre- and post-test changes of student understanding, both quantitative, and qualitative. In the case of astronomy we have been able to show that on an internationally recognized test, the improvement from pre- to post-test is

consistently twice that reported by previously reported interventions (Hubber, 2010).

This list illustrates a range of forms of evidence associated with student and teacher classroom behaviors that indicate high-level performance, yet do not directly translate into 'outcomes' that students and teachers can necessarily demonstrate in formal terms beyond the context in which these behaviors were evidenced. A distinction needs to be made between processes that occur during an innovation that are identified as high level by virtue of conforming to recognized indicators of quality, and products that may be knowledge, or capabilities, that are demonstrably a longer term outcome demonstrable in an abstracted form beyond the particular context in which these were learnt, or guided to perform. Yet given the highly contextual nature of the way high-level processes may be enacted, I argue we need to find ways to rigorously evaluate such embedded performances in situ.

There have been a number of pressures that have meant the approach has not been subjected to a formal experimental study. First, concerning the research and development cycle, the funding has supported researchers working with teachers to develop and refine the approach and not an evaluation involving control groups implied by a scaled up version of a developed and packaged product. Second, there are practical and ethical issues surrounding such a methodology, for instance in subjecting control classes to assessment on outcomes specific to the innovation, which may be inappropriate for traditional teaching. Thus, in this project, again we need ways of describing a range of forms of evidence within a rigorous evaluation process that does not require a formal experimental design.

Case 3: Evaluation of the model and impact of the Scientists and Mathematicians in Schools (SMiS) program

In this study of STEM professionals partnered with teachers in a formal arrangement, evidence was generated by a number of processes including a) a survey of teachers and STEM professionals, b) interviews with members of the SMiS team running the program, and c) development of case studies of partnerships using interviews and focus groups.

The evaluation followed the logic model developed by the Kellogg Foundation (2004), shown in Figure1.In this case the outcomes and impacts are woven into the evaluation narrative, which also takes account of the program activities and outputs leading to outcomes and wider impacts. The model allows for a more flexible and inclusive consideration of program worth beyond simply formally measured outcomes, and the consideration of wider impacts beyond the immediate reach of the program itself.

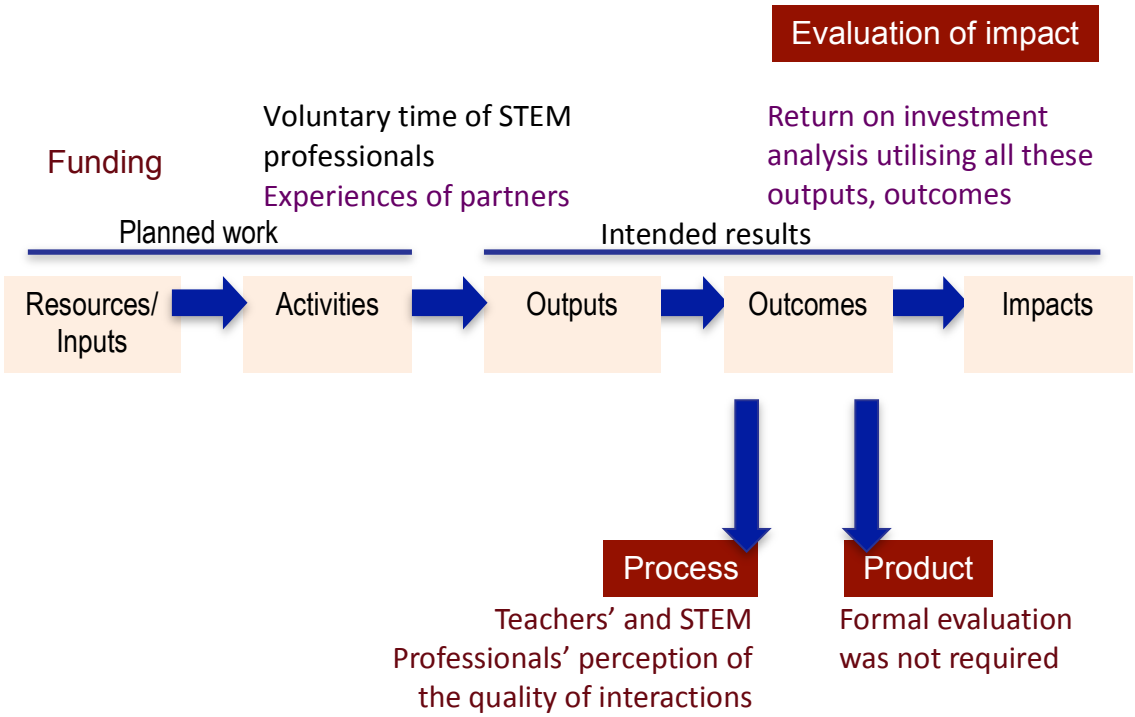

Figure 1. The logic model driving the SMiS evaluation (based on that of the Kellogg Foundation)

The categories include:

- Inputs: monetary input for program support structures. These can include reputational resources of the CSIRO organization running the program.
- Activities: Activities of the CSIRO SMiS team in arranging, setting up and supporting partnerships and networks, monitored using survey and interview tools. Activities in schools which were monitored using survey tools and which could be identified as focused on significant curriculum innovation, and case studies.
- Outputs: Time commitment of the STEM professionals, which amounted to a 3-fold return of investment taking account of STEM professional time leveraged by the project input and partnership activities in schools. The time spent by students and teachers in contact with STEM professionals, which could be quantified by considering the cost of equivalent activities.
- Outcomes: Changes in student, teacher and STEM professionals. Monitored by survey of teacher and STEM professional perceptions, and interviews.
- Impact: Evaluated by pulling together the variety of forms of evidence of activities, outputs, and outcomes and developing a narrative argument that included consideration of the context of the program within the Australian education and employment scene.

In the evaluation, the findings are spread across a variety of forms of evidence that include not only a) product outcomes as identified by participants, but also b) the nature of activities and outputs, judged according to where these sit within established quality principles established through the literature, and c) projected impact that is established through the setting of the program within the broader sweep of educational provision, and within an established literature on partnership activities of this kind.

Discussion/conclusion

I have argued in this paper that an exclusive focus on experimental methods fails to recognize the complexity at all levels of education systems, and drives evaluations of innovation towards consideration of a narrow range of outcomes only. These outcomes tend to be short term, product outcomes that fail to recognize the complexity of the more important aims of education systems, as illustrated by the Melbourne Convention, and the complexity of process outcomes that are inevitably contextual and contingent and hard to capture on standardized measures.

We need to acknowledge evaluation approaches that recognize a wider range of performances in context that can be shown to contribute to the development of important education aims, and which incorporate a variety of methodologies each of which has its established standards of rigor. I speculate, with respect to the question of rigor, that there are particular research programs that will contribute towards establishing more robust responses to this complexity. One is the development of a research literature that identifies features of quality performance on higher order learning processes, under contingent conditions, which can be used as benchmarks of quality identification. The second is the development of processes and standards of narrative investigation and reporting that can bring together such literature with rich descriptions of the experience of participants in innovation settings, together with sophisticated analyses of the setting of the innovation within educational, and broader social frames. The third is a program of robust critique of the current neo-liberal fascination with measurable outcomes, which demonstrates the longer term, demeaning effect of such reductive versions of teaching and learning.

References

Australian Curriculum, Assessment and Reporting Authority (ACARA) (N.D.). Science—Foundation to Year 12. Accessed 24 Dec 2015 at http://www.acara.edu.au/curriculum/learning_areas/science.html

Australian Education Ministers (2008). Melbourne Declaration on Educational Goals for Young Australians. Accessed Dec 24 2015 at: http://www.curriculum.edu.au/verve/_resources/National_Declaration_on_the_Educational_Goals_for_Young_Australians.pdf

Cohen, L., Manion, E., & Morrison, K. (2000). Research Methods in Education. (5th ed.) London: Routledge Falmer.

Hempenstall, K. (2006). What does evidence-based practice in education mean? Australian Journal of Learning Disabilities, 11(2), 83-92.

Kellogg Foundation. (2004). *Logic Model Development Guide*. Michigan: W.K. Kellogg Foundation.

Lawrenz, F., & Huffman, D. (2006). Methodological pluralism: The gold standard of STEM evaluation. New directions for evaluation, no. 109. Wiley.

Maxwell, J. (2004). Causal Explanation, Qualitative Research and Scientific Inquiry in Education. EducationaI Researcher, 33(2), 3-11.

Shelley, M. C. II, Yore, L.D., & Hand, B. (Eds.), (2009) Quality research in literacy and science education: International perspectives and gold standards. Dordrecht, The Netherlands, Springer.

Tytler, R. (2007). School Innovation in Science: A model for supporting school and teacher development. *Research in Science Education*. 37(2), 189–216.

Tytler, R. (2009a) Longitudinal Studies into Science Learning—Methodological Issues. In M. C. Shelley II, L. D. Yore, & B. Hand (Eds.), Quality research in literacy and science education: International perspectives and gold standards (pp. 83-106). Dordrecht, The Netherlands, Springer.

Tytler, R. (2009b). School Innovation in Science: Improving science teaching and learning in Australian schools. *International Journal of Science Education*, 31(13), 1777-1809.

Tytler, R., Prain, V., Hubber, P., & Waldrip, B. (Eds.). (2013). *Constructing representations to learn in science*. Rotterdam, The Netherlands: Sense Publishers.