University of the West of England
BRISTOL

Faculty of Business and Law

How should economics curricula be evaluated?

Andrew Mearman
Department of Accounting, Economics and Finance
University of the West of England, Bristol, UK

Economics Working Paper Series
1306

How should economics curricula be evaluated? [1]

Abstract: This paper explores the evaluation of economics curricula. It argues that the dominant approach in economics education, experimentalism, has serious limitations which render it an unsuitable evaluation method in some cases. The arguments against experimentalism are practical, ethical and also rest on a view of the world as a complex, open system in which contexts are unique and generalised regularities are unlikely. In such an environment, as often found in educational contexts, alternative methods are advisable, at least as part of a suite of approaches in a realistic, case-based, mixed-methods approach to evaluation. Thus, economics curricula should be evaluated using a method or set of methods most appropriate to the particular object case. As such, there is no single answer to the question posed.

JEL classification: A20, A22, B4, B5, C80, C9

1. Introduction

This paper addresses the question of how to evaluate economics curricula. Evaluation is central to educational practice and improvement. As a reflective practitioner, the educator will try different combinations of content and delivery, in an effort to achieve their particular goals. Increasingly, also, there are demands from the educational literature (for example, Hargreaves, 1997, 1999; Oakley, 2007) and from the literature on economics education in particular (Davies and Guest, 2010) that claims of effectiveness of teaching innovations are supported by field evidence. This paper is part of a response to these calls. Its argument draws on the wider literature on education and research methods.

The paper argues that the dominant approach in economics education, experimentalism, has serious limitations, such that it is unsuitable as the evaluation method in many cases. The arguments against experimentalism are practical and ethical, but also ontological: they rest on a view of educational

[1] A version of this paper was a plenary address to the conference of the International Confederation of Associations for Pluralism in Economics (ICAPE), University of Massachusetts-Amherst, on 11 November, 2011, and a subsequent version was delivered as a UWE staff seminar on 15 March, 2012; I should like to thank participants of both for their comments. A subsequent version was published as a UWE Discussion paper 12/03. Thanks to Peter Davies, Tony Flegg, Mary Hedges and Don Webber for comments.

contexts as complex, open systems in which contexts are unique and generalised regularities are unlikely. In this environment, alternative methods are advisable, at least as part of a suite of approaches in a realistic, case-based, mixed-methods approach to evaluation. Thus, economics curricula should be evaluated using a method or set of methods most appropriate to the particular object case. As such, there is no single answer to the question posed in the paper's title. The paper is structured according to that argument.

2. Context

Teaching practice may be said to develop via trial and error. Teachers experience the need to innovate, often in response to an evaluation that prior practice did not 'work' as well as hoped. Such an evaluation might be based on student feedback or achievement – informal or formal – or simply on the reflections of the teacher that students were not *inter alia* engaged, learning effectively and/or attending. The reflective teacher also contemplates why something has been (in)effective. That teacher may try to attach their practice to some educational (or other) theory. Finally, they might try their innovation on another group of students. In some cases, they try to evaluate their innovation formally, via a research project.

The economics discipline has recently experienced a marked increase in publications on developing effective teaching and its evaluation. Possible drivers for this are, *inter alia*, a greater intrinsic interest in teaching and its effectiveness; a global recruitment crisis of economics students in the 1980s and 1990s; a growingly competitive global marketplace for students; greater scrutiny and quality assurance from layers of governance, governmental and institutional; and an increasing focus on achieving high scores in student experience or satisfaction surveys published in league tables.

More broadly, there have been specific calls that evaluation of curricula and other areas of innovation should be grounded in systematically conducted research (Davies and Guest, 2010). Some calls have been for educational reform and policy to be *evidence-based* (Hargreaves, 1997) or *evidence-informed* (Hargreaves, 1999); for counterarguments, see Elliott (2007), Gage (2007) and Bassey (2007). Overall, teachers are asking themselves *some variant* of the question 'what works'?

Assuming educators ask 'what works?' they must consider two things: their criteria, and their evaluation processes. This paper will focus on the latter, but it is essential to consider the former, as there is a link between the educator's aims of education, and thus their criteria, and the tools they choose to evaluate their practice. These aims will vary between educators and may vary in terms of many dimensions.

Some aims are externally imposed, for example by the educator's institution or profession, or by the state. Individual institutions may demand that courses have pass rates or mean marks which exceed a threshold. Academics usually believe that disciplines have core understanding that students should demonstrate. Some disciplines – although not economics – have strong professional body requirements to deliver specific skills or other content. State bodies such as the UK Quality Assurance Agency place some conditions on practice. Additionally, all instructors – or in a programme, each team of instructors – have their own aims, even if they do not know this.

Therefore, curricula have plural aims. Clearly with multiple aims it is possible that an innovation could lead to one aim being achieved more successfully, while another is achieved less well. Thus: curricula could 'work' in numerous ways, only some of which are consistent with each other. That suggests that there are no universal standards by which teaching efficacy can be judged.

Further, some aims may be difficult to evaluate – they are deeper, generative; not superficial. Clarke and Mearman (2004) borrow the analytic separation of aims into liberal and instrumental[2]. Liberal education aims to create indeterminate capacities within students: opening the student mind via critical thinking, independent analysis and comparative thinking, leading to the exercise of autonomous judgement. Instrumental aims tend to focus on more concrete outcomes: being able to understand x, relate theory y, solve specific problem z. They may go beyond this to aim at employability, or control. Instrumental aims tend to be more short term and concrete; and thus are easier to evaluate. These two positions are poles, and not mutually exclusive: Guarasci (2001) and colleagues perform the 'practical liberal arts' in which the two aims are achieved through practical engagement with real situations.

How do teachers reflect on and improve their practice? One might expect them to engage in ongoing reflection, using their own judgement and intuition. Indeed, for researchers adopting one stance, this is the essence of good teaching practice: as the teacher has unique goals, teaching in a unique context, generating outcomes with unique meanings, personal reflection is the only method appropriate.

However, teaching innovations or curricula may be evaluated via a formal research process. This research may be conducted by the practitioners themselves, or by outside researchers. Whoever does the research, it involves a series of choices on the evaluative framework, data types and methods of analysis. There is a range of options. In evaluation literature there are *inter alia*, experimental, pragmatic and realistic approaches (Pawson and Tilley, 1997). The first is largely quantitative, the third largely qualitative, and the second a mixture.

Accordingly, in the educational literature there is a paradigm struggle between qualitative and quantitative methodologies (Gage, 2007). Qualitative researchers argue that their context-specific methods are more appropriate to the nature of educational processes. They assert the particularity of case and of the practitioners' aims, and therefore the need to build this into the research process. However, their opponents charge them with radical particularism. Furthermore, advocates of quantitative research criticise qualitative methods as imprecise, and subject to sample and crucially researcher bias: they tend to argue for the possibility of greater objectivity, replicability, validity and reliability.

3. Experimental methods

Outline

A common approach (see, for example, Opie, 2004; Morrison, 2009) is to evaluate curricular reforms or teaching techniques experimentally. Indeed, in economics education, experimental methods dominate. This fact partly reflects the utility of experiments in evaluating teaching (for example, see Dickie, 2006; Contreras, et al. 2012), and as a teaching tool (for example, see Watts and Guest, 2010, and associated papers). It also reflects a more general growth in experimental activity in economics. Furthermore, as Pawson and Tilley (1997) note, experimental design is often favoured by proponents of evidence-based policy (EBP): Martinson (1974) only included experimental studies in his seminal systematic review of prison reform programmes.

'Pure' experimental design in education and other social sciences is modelled on the OXO formula, in which an experimental group and a control group are tested (the Os) before and after an

[2] Hargreaves (1999) offers a similar dichotomy between engineering and enlightenment education.

intervention (X) in a process which is only applied to an experimental group (Campbell and Stanley, 1963).[3] The mechanics of these tests are simple: two groups would be taught (preferably nominally the same course); one group (the experimental) would be exposed to some teaching innovation. The other group (the control) would not be exposed to the innovation and be taught as before. At the end of the course, both groups would be tested.

Ideally, the two groups are identical in terms of key characteristics: their members are 'matched pairs'. In a pure experiment, the scientist selects members randomly. Furthermore, the scientist creates a controlled environment such that any difference between the two groups (post-intervention) is directly attributable to the intervention. Hence, these are known as Random Controlled Trials (RCTs). Experiments can take many forms, including laboratory tests of various types.

The assessment tool chosen would depend on the goals of instructors and their predispositions to particular data types. So, for example, one option is simply to use a multiple-choice test to assess accuracy of knowledge of theoretical material. If the innovation is effective, one might predict that the average exam score of the experimental group would be higher than the control group.

Whatever measure is used, statistical tests are often employed to assess difference (Opie, 2004). In regression analysis the test scores are the dependent variable and the intervention (or absence of) is an independent variable (Opie). Researchers might infer causality from these tests. If random sampling has taken place, this causal inference is often extended to the population. Such claims are stronger when supported by meta-analysis, and by replication (Pawson and Tilley, 1997). Such evaluations are often used to support policy-based reform of public sector practices.

Critique

RCTs have several claimed benefits. Above all, experiments represent the 'scientific method'. Experiments are claimed to be replicable (and thus verifiable), systematic, and hence more likely to generate reliable and valid results. Thus, they are also claimed to offer greater scope for generalisation to other cases; and therefore greater impact (Oakley; Bolgar, 1965, cited in Schofield, 2007). Many authors reject qualitative and ethnographic methods as having bias. Further, Oakley (2007) claims that, in comparison to methods based on personal reflection and judgement of the researcher, experimental methods are transparent: those affected by the research (principally students and parents) can observe its processes. This is particularly true when the research is done independently (Hargreaves, 1997).

RCTs have other attractive characteristics. They appear particularly suited to the assessment of short-run interventions and effects, which are attractive to educational practitioners who usually face the constraint of limited (relatively short) access to student groups. Further, because short-run aims tend to be relatively concrete and instrumental (for example, more effective learning of specific concepts) they are suited to experimental evaluation. Additionally, there is a widespread emphasis on summative assessment of learning at frequent intervals, again suggesting a role for short-run

[3] The size of the intervention may also be varied across the experimental group (Morrison, 2009). So, a group of people may take different strength variants of a medicine, to check how much is needed to cure the particular ill. Unfortunately, it would be extremely difficult to design an experiment within an institution which allowed some students exposure to different degrees of innovation. The ethical problems discussed below would be considerable.

evaluation of students and thereby of teaching innovations. This emphasis may reflect what Elliott (2007) calls Outcomes-Based Education (OBE).

Furthermore, RCTs reflect prevailing characteristics in economics. Proponents of statistical analysis advocate RCTs, because they appear to provide robust quantitative data amenable to analysis by sophisticated statistical techniques. Also, regression equations are (erroneously) compared to experiments: the effect of each cause is measured assuming others are constant; and the equation error term is held to be analogous to closing the laboratory door[4]. A recent example is the development of difference in differences models, which are applied to natural experiments (see, for example Bonesrønning and Opstad, 2012). All of these models reflect the control model (Elliott, 2007); i.e. assume that the scientist/educator has control over their environment. To some extent this control generates predictable outcomes. These assumed properties reflect further a shared belief in a regular world, or *closed-systems ontology* (Lawson, 1997).

However, good experimenters know that the process of conducting RCTs involves several challenges. Crucially, we cannot avoid the fundamental problem of causal inference (Holland, 1986): that one person cannot be in the control group and the experimental group at the same time. Even if the experiment could be designed so that the subject(s) were in the control group and *then* the experimental group (or vice versa), causal transience (Holland) may occur: it is possible that being in (out) of the control group has an effect which is then felt when the subject is out (in) the experimental group (see Morrison, 2009: 159).

It may not be possible to set up groups of matched pairs, or even to set up a robust control group. For instance, certain types of people will opt into the experimental group – they might inherently favour non-standard curricula – leading to sample-selection bias. To counteract this 'volunteer effect', the experimenter may split up the volunteers into an experimental and a control group (Pawson and Tilley, 1997). However, students who had signed up for a particular type of course would likely recognise they weren't receiving that, and the experiment would break down.

One can avoid volunteer effects by evaluating innovations only on core courses. However, where one group has access to a potentially advantageous intervention, yet another does not, serious ethical issues are raised (Sikes, 2004: 25). Experimenters are aware of these problems, without clear resolution. Campbell (1969) offers an attractive vision of an experimenting society in which there is a social contract whereby experimentees are compensated but also where the benefits of experimentation are shared, and thus the risks too. However, that view might be considered naïve. Further, Oakley's (2007) critique of qualitative methods would apply here: experimental subjects are effectively compelled to be so.

Given these practical challenges to pure experiments, alternative designs are tried. Two-group tests without random selection are possible; although this raises selection-bias issues again. One main problem with matched pairs design is that the paired subjects have to be identical in terms of all other variables (or as many as possible). This is challenging, not least logistically, as it would involve a huge data collection process prior to the experiment (Borg and Gall, 1996). In these cases, a non-equivalent pairs, quasi-experimental design may be chosen. As Opie (2004) notes, this is often the only available option in educational research. Lim (1998) used such a design to test the effectiveness of WinEcon. He did some pre-testing of aptitude, and of demographics, to use as control variables in explaining the post-intervention exam scores. Lim's study is an example of a field experiment.

[4] This is a false analogy because in regressions, experimental conditions are imposed on data created under uncontrolled conditions.

Field experiments (see Barankay, et al. 2013) seek to "achieve sufficient *control* to make the basic causal inference secure" (Pawson and Tilley, 1997: 6, emphasis in original). However, as Morrison (2009: 159) notes, field experiments suffer the criticism that the participants are too different to make meaningful comparison possible.[5] Partly as a response to that criticism, one-group experiments are done, in which the experimental group is examined pre- and post-test to see if the intervention has made any difference. These designs are clearly inferior to pure experiments but often are all that is feasible in many educational institutions. Indeed, given the nature of the research into curricula, the barriers to conducting even one-group evaluations are considerable.

In addition to these practical criticisms of experiments, more fundamental objections are raised. For example, experiments may generate black box theories. For Pawson and Tilley (1997: xv) "…experimentalists have pursued too single-mindedly the question of whether a program works at the expense of knowing why it works". Thus, experimental results merely generate statistical generalisations (Elliott, 2007). Similarly, for Morrison (2009: 157) "…'what works' may fail to address causation at all; causation is about 'how something works', not only 'what happens'." Moreover, as discussed, "…judging 'what works' is a matter of values and not only of performance" (Morrison: 172).

Furthermore, an experimental approach can lead to biased analysis – experimenters may only look for "confirming evidence, and… not seek, or find, rival, alternative explanations, nor … seek falsification criteria" (Morrison, 146). Experimentation can close off possibilities of other hypotheses emerging (Morrison, 155). A further problem with Martinson's approach is that, if experimental methods are flawed in essence, then only comparing experimental results with others is in turn flawed: it would be advisable to compare them with results of different types (see Siakantaris, 2000).

A further criticism of RCTs is that experimenters may seek universally applicable results, or interventions which are generally shown to be effective and beneficial. In this mindset, interventions are expected to work in all cases; or not at all. Hence, Martinson (1974) reached the conclusion that 'nothing works' because he could not find universally applicable results from the experiments he examined. However, "…there is no approach that can ever guarantee universal learning success, however success is defined" (Hodkinson et al. 2007: 37).

Even where one seeks to find interventions which are relatively likely to be more successful, RCTs may not offer clear guidance. Experimental results often display variety: "the same program will thus work in quite different ways for different subjects and … the experimental method is simply not designed to appreciate such subtleties" (Pawson and Tilley, 1997: 31). This problem is compounded by statistical tests which measure average effects.

Another criticism is that the treatment of historical time is inadequate in experimental approaches. Causal relationships may suffer temporal complication: for example, smokers do not immediately get lung cancer (Morrison, 2009). Short-term educational interventions – courses often only run for a term – may have few immediate, discernable effects (Morrison, 2009: 141). It is unlikely that the liberal goal of opening student minds would be observable in the short term. Furthermore, as Opie (2004) points out, the longer the experiment, the more difficult it is to remove unwanted confounding effects of other factors: not all smokers contract lung cancer.

[5] Even more *impure* are natural experiments in which no control over events is feasible, but where a policy intervention has occurred and one can try to study its effects.

This somewhat lengthy critique of RCTs is defensible given the dominance of that method in economics education. Collectively, these criticisms suggest that, experimental methods have serious flaws. Significantly, many of them (for example, concerning bias or validity) are the same as those identified in qualitative research. Collectively they suggest strongly that it would be wrong if experiments were the only acceptable, or default method. Hammersley (2007: 63) warns that advocates of independent, experimental research "must not ... exaggerate the prospects of success at the expense of other kinds of research". He echoes strongly Lawson's (1997, *et passim*) arguments that economists should not insist on mathematical and/or statistical modelling, because they are only the most appropriate approach, or indeed appropriate at all, in specific circumstances.

However it seems that experimental methods are best – perhaps only – employed under certain conditions: where it is possible to create groups which are (albeit imperfectly) separable; where doing so does not create ethical difficulties; where the intervention has a short-run nature; where the goals of the intervention are short-run and instrumentalist. In other cases, it might be better to use other methods. There is at least a case for greater flexibility in experimental design, and combining experimental results with those of different types.

4. An alternative approach

To reprise: in the case of a typical experimental study, the researcher typically sets up an experimental group; has short-run and instrumental goals; aims to measure success often using test scores; and hopes to generalise these results beyond their own. However, the researcher may decide that in their particular case, experimental methods would be problematic. If so, what other options are available to them? There are many other options open to the researcher, either as alternatives or complements to experiments. This section discusses these options. First though, it is necessary to clarify some key ontological points.

Ontology

Underpinning the above critique of experimentalism are fundamental criticisms of its implicit view of the world. The experimental method works on the premise that the scientist can control a situation within an environment to such an extent that they can trigger a cause, having a discernible effect, rendering other causes negligible; and that they can do this a required number of times, leading to a consistent finding.

In contrast, many authors argue that the world comprises complex, open systems, exhibiting equifinality, emergence and internal relations. In such a world, causes may not be separable. Thus, the neat separation of cause and context presupposed by experiments may be infeasible. As Morrison (2009: 107) says, "[a]ny cause or intervention is embedded in a web of other causes, contexts, conditions, circumstances and effects, and these can exert a mediating and altering influence between the cause and its effect".

It seems that educational environments are of this type: they include "teaching, teachers, learners, learning situations, and wider historical, economic, social and political influences" (Postlethwaite, 2007: 161). For James et al. (2007: 11) "…teaching and learning cannot be decontextualized from broader social, economic and political forces, both current and historic". Hence, life outside college cannot be separated from life in college, learning sites cannot be separated, the individual and the social may not be separable (Hodkinson et al. 2007). More prosaically, control and experimental groups may not be separable because of social media (Morrison, 2009: 156).

Any effects of curricula are complexly determined. They include understanding, knowledge, criticality, engagement and the development of other generic skills such as research, use of literature, data analysis, etc. These factors are interdependent. Learning may be said to presuppose some extent of engagement, which is itself multiply determined, by the approach and abilities of the teacher, the curriculum, the nature of the learning environment, and other factors such as students' backgrounds, group dynamics and the like (Zepke, 2011). If we accept that (particularly, liberal) educational benefits may take some time to accrue (because they are concerned with deep cognitive and attitudinal tendencies), then the complexity increases, because factors such as the general cognitive capacities of the student, the nature of their external environment, their teachers and the subjects they take may be changing.

Crucially, the dispositions and responses of people are key in determining whether an educational intervention works. Learning involves a conscious effort, and this may occur because of pre-existing subconscious predispositions, which are themselves changed in the process of learning (Hodkinson et al. 2007: 35). Ultimately, learning is something people do, not something that is done to them (Hodkinson et al.: 34). As Pawson and Tilley (1997: 33) argue, participants take a cognitive leap, necessary for the intervention to work. To be successful, programmes require volition; co-operation (or mindset) may be the crucial factor which determines success (Pawson and Tilley: 36).

By contrast, experiments must design out these factors. Thus, in educational fora, the dispositions of teachers (James et al. 2007: 12) or students may be assumed away. In trying to avoid the volunteer effect, volition is designed out (Pawson and Tilley, 1997). RCTs necessarily assume that the meanings held by teachers and students in their experimental site(s) are the same: this is a vital, but perhaps too bold an assumption. This might be summarised as the experimental method designing out individual people from the process (Elliott, 2007). In such a world, "personality, creativity and so on are irrelevant" (Opie, 2004: 87).

Overall, "…striving to control the influence of extraneous factors by random assignation of participants to control and experimental groups … is ill-judged, as this prevents researchers from identifying those very conditions that might be contributing to the success or failure of a programme… i.e., precisely the sort of information that might be useful to policy makers" (Morrison, 2009: 154-5).

A realistic, case-based approach

For the practitioner sceptical of experimental methods, particularly if informed by the above ontological concerns, another framework is available which embraces the ontology of complex, open systems. It is found in three literatures: realistic evaluation (Pawson and Tilley, 1997), case-based methods (Byrne and Ragin, 2009), and pluralist or mixed-methods research (Downward and Mearman, 2007). Collectively, these literatures stress that research should be based in real cases, utilise methods that take context seriously, and be nuanced in their claims made. They all embrace the critiques of experimentalism discussed above.

Pawson and Tilley (1997) discuss various evaluation approaches. They argue for realistic evaluation, which employs no one standard "formula" (Pawson and Tilley: xv), but like pragmatist evaluation, to some extent is characterised by a plurality of techniques and "on the craft skills of the researcher" (Pawson and Tilley: 15). However, Pawson and Tilley favour realistic evaluation because it is wedded to the well-developed ontological position described above. Accordingly, Pawson and Tilley in particular are quite sceptical of universal claims of efficacy for interventions. In turn they reject the model of evaluation which rests on regular successions of events, control and the strict relation

between inputs and outputs which characterises Evidence-Based Policy (and by implication, Outcomes-Based Education) as discussed above. Instead, they advocate individual case studies. That connects well with the case-based methods literature.

Case-based methods attach primacy to the case. There is much debate on the definition of a case (see Ragin and Becker, 1992). To some extent this question remains unresolved. Nonetheless, we can say that a case can be at a number of levels (nation, region, organisation, department, individual). It is "a phenomenon of some sort occurring in a bounded context" (Punch, 2009: 119). A case involves a process as well as an outcome (Mjøset, 2009). Cases are unique and thus special attention must be paid to the individual history and context of each one. Byrne (2009) notes that one definition of a case is as a complex system.

From our discussion of learning environments above, they would seem to fit the definition of cases. For Bassey (2007) education is the science of the singular. A case study is an in-depth investigation of a case. As Goode and Hatt (1952, cited in Punch, 2009) say, a case study is not a particular technique: hence an experimental approach could be part of/the case study.

Case-based methods are claimed to have many advantages. They explore in detail the specific facets of a case, often including explanations of the observed patterns of behaviour or learning outcomes. To do this they must elaborate the specific nature of the individuals and their learning environment. As discussed, one of the criticisms of experimental methods in educational research is that they may ignore the specific natures of the teacher, the learning environment, and the students. While experimental approaches do require heterogeneity within their sample (in order to show effect over a variety of subjects), they tend to assume that what works in one case will do so in all. Case-based researchers would argue that this assumes too much. In contrast, case-based methods may be able to explain the failure of experimental methods to generate clear results or achieve meaningful interventions. They counteract the failure of the experiment to capture sufficiently details of the case. Further, they might show when experiments might be used.

For all of these reasons, case studies are well established in educational research. Significantly, relating to Hargreaves' (1997) arguments for (largely) experimental evidence-based policy, case studies are of course highly significant in medical research (Punch, 2009: 122).

As case-based research does not dictate any particular methods of analysis, it is difficult to either advocate one set, or criticise case-based methods for attachment to that set. However, case-based methods are criticised: often because they typically deploy qualitative methods, which are alleged to lack reliability and internal validity and are dogged by researcher bias.

Connected to this set of criticisms is a principal objection to case studies: that cases cannot be generalised (Bolgar, 1965). Further, an interpretivist researcher might claim that cases *should* not be generalised because this robs the case of its uniqueness. In comparing cases, some crucial, relevant detail may well be lost. Not least, the perspective of the researcher – which is regarded as crucial – is lost. Thus, replication of cases (or experiments) is impossible and undesirable. This position reflects an ideographic (as opposed to nomothetic) approach to research, in which inference beyond the case is unattainable. Guba and Lincoln (1982) argue this, because each case is time- and context-dependent. Indeed, as Schofield (2007) notes, qualitative case studies may be designed to capture change over time, rather than a short-run snap shot.

However, recently there have been attempts to generalise using cases, often with qualitative data (Schofield, 2007). For example, Cronbach (1982) argues that one case study provides a working

hypothesis for another. Recent case-based methodology reflects an established contextualist tradition which bridges the disparate aims for knowledge of completely unique cases, and of context-free general theories (Harvey, 2009; Mjøset, 2009; Byrne, 2009). In this approach, while generalisation is difficult because of causal and conceptual heterogeneity (Mahoney and Terrie, 2009), and should be limited, it is recognised that cases are *cases of something* (Goertz and Mahoney, 2009), and that as such it ought to be possible to build up knowledge by comparing similar cases, using historical case knowledge (Rihoux and Lobe, 2009; Burawoy, 1998). There is some support from the literature on grounded theory for this position (Mjøset).

Overall, rather than aim for universal or general findings, these research approaches aim for relatability – or translatability, fittingness (see Schofield, 2007), or comparability (Goetz and LeCompte, 1984) – the ability to say something about other cases, where possible. Schofield also discusses how relatability can be increased. She suggests that by choosing 'typical' sites, one can be more confident that findings applying there may also hold in other sites. The notion of a typical site is problematic, as it would probably require analytic induction to establish it. Further, thick descriptions of each case would still be required in order to assess its typicality.

Schofield (2007) also offers support for the notion of comparing dissimilar cases. She recommends using multiple sites partly for this reason. More broadly, there is support some authors in the case-based methods literature are interpretable as favouring a form of inference similar to negative analogy (Keynes), i.e. finding similar conclusions in dissimilar contexts suggests that the conclusion has greater weight. Again, the analysis would require thick descriptions of each case and their comparative analysis. This is often time consuming and requires access to multiple sites: that may be logistically impossible.

The broad educational literature reflects many of these themes. We can recognise the notion that "….a case is an outcome preceded by a process that unfolds in time" (Mjøset, 2009: 47) in the myriad case studies in the literature, which present detailed, specific institutional analysis. James et al. (2007) is one example, which deploys many of the principles outlined above. Their methodology deploys a number of case studies, all of which recognise the specificity of each, yet are linked as cases of learning cultures. However, as discussed, realistic evaluation designs discussed here aim only for relatability.

Mixed-methods design

Thus far, we have discussed a realistic evaluation framework based around case-based research, founded on a particular ontology. However, specific methods have not been considered. Case studies are not a method and may deploy a range of methods. Often researchers choose qualitative tools, which encompass a diverse set of research tools such as personal documents, semi- or un-structured interviews, critical incident analysis (Tripp, 1993), observation and other ethnographic techniques; these are usually geared towards providing detailed and deep understanding of specific contexts; and they are often rooted in philosophical approaches to research which reject the alleged positivism of quantitative or experimental work, such as action research, grounded theory or interpretivism. These approaches are well established in educational research (see, for example, Punch, 2009).

This paper advocates mixed-methods design. The often strict distinction drawn between qualitative and quantitative research (also often conflated with experimentalism) is unsustainable. For example, all quantitative data assumes qualitative invariance of the object (see Downward and Mearman, 2007). Similarly, it is false to strictly separate experimental and qualitative methods, since there is

considerable variation within experimental design, and within that much scope for qualitative elements. Further, much (perhaps most) experimental work in economics education is done by practitioners, who are hoping to study change as they create it: this is the essence of action research.

Mixed-methods research is a well-established tradition (see Creswell and Plano Clark, 2011). Feminist Economics has made strong arguments for mixing methods. Downward and Mearman (2007) suggest ontological arguments for mixing methods: complex objects demand a combination of data types and analytical techniques, with the exact combination being driven by the nature of the question(s) being asked and the object(s) under study. Quantitative methods can offer a broad analysis of the object under study but not the depth of qualitative, and therefore both are necessary; further, all methods are inherently fallible and therefore that no single method of assessing achievement of learning outcomes could be comprehensively successful. Mixed-methods research rejects the notion that any one method (for example, experiments) should be *insisted* on.

There is no algorithm for arriving at a combination of methods. However, following Denzin (1970), we would expect to 'triangulate' different data types, investigators (or sites), theories and methods of data collection and analysis in order to arrive at a more rounded and more convincing conclusion. Significantly, "…realistic evaluation can utilise a range of research designs and so can be quantitative or qualitative, action- or outcome-oriented, contemporaneous or retroactive…" (Pawson and Tilley, 1997: 182).

Mixed-methods approaches are well established in the educational literature. This is evident via the wide range of data used; and via their *interaction*. First, educational research utilises quantitative data and analysis, for instance through large-scale survey (using questionnaires) data, generating different data types, including bi- and multi-variate response data (Morrison, 2009) analysed using a range of statistical tests. However, the literature also deploys various qualitative data collection and analysis methods. For example, qualitative semi-structured interviews, focus groups, student learning journals, physical institutional artefacts, tutor professional journals, small group discussions and other ethnographic methods are all evident (Opie, 2004; Soh, 2001; Hodkinson, et al., 2007; Postlethwaite, 2007; Morrison). These, and other methods discussed above, fit the longitudinal approach necessary to assess the effects of educational curricula, given the time lags and the nature of the developments involved.

Second, educational research embraces the benefits of consulting different types of people in different places at different times, because "different students, different tutors, different college managers, employers, parents, and policy-makers, will have differing views about the outcomes that are desired…" (Hodkinson et al. 2007: 36). Postlethwaite (2007) describes how, in one large-scale project, a range of data collection and analysis techniques is employed at a number of different learning sites; for instance, covering different types of institution and types of course (see James et al. 2007; Hodkinson et al.: 25). Postlethwaite also stresses the utility of student perspectives data (171). Morrison (2009) demonstrates a range of reporting methods: another example of mixing.

Third, and crucially: "The more the researcher wishes to understand causal process, the more methods in combination are useful, each with their own time frames and timing of data collection" (Morrison, 2009: 169). James et al. (2007) deploy a combination of qualitative and quantitative techniques. Their work exemplifies a quanQUAL (Creswell and Plano Clark, 2011) approach: quantitative data are used extensively but qualitative research is emphasised. Similarly, Postlethwaite (2007: 170) describes how a survey would generate quantitative results, prompting further qualitative investigation of the statistically significant ones. Moroever, Postlethwaite (174) envisages quantitative and qualitative findings placing a check on each other: "We analysed the

qualitative and quantitative data separately then compared insights, finding enough synergy to suggest that findings were not artefacts of one or other method".

The logic of mixing also applies to metric used. Given that in curricula there tend to be multiple goals, a single metric is unlikely to be appropriate. Let us recall the liberal goals discussed above: analytical, comparative and critical thought. Critical thinking may be assessed through more detailed questions or scenarios, such as research essays or long written answers in exams. At this point the assessor's judgement evaluates whether and to what extent the student has demonstrated critical thinking, made effective criticisms, applied the theoretical tools appropriately, or reached a sensible conclusion.

Alternatively, there is a range of standardised tests which aim to assess critical thinking skills. Among these are the California Critical Thinking Skills test. However, these are all subject to criticism that they reward memory, or training, or those suited to the specific type of test. Cottrell's (2011) critical-thinking exercises are designed to be developmental and would be inappropriate as 'before-and-after' tests of critical thinking. Kirby et al. (2002) use Bateman and Crant's shortened Personal Proactivity Scale to assess students' ability to use initiative, which could indicate autonomy. However, like the others, these tests are highly time consuming and logistically challenging. Unless they replace the main assessment tool, they may also cause assessment fatigue in students.

In the nascent literature on the efficacy of pluralist economic curricula, a large range of assessment tools have been deployed (O'Donnell, 2009; Earl, 2000; Resnick and Wolff, 2011; McIntyre and van Horn, 2011; Lapidus, 2011; Barone, 2011, 2012; Amin and Haneef, 2011; Stilwell, 2011). Specific assessment tools deployed include student essays and student evaluations; interpretive discussion; thought papers; and factual checks. In order to try to capture the long-term effects of curricula, some researchers favour learning diaries, which can record a student's thought processes longitudinally through a module of study. In the case of the learning diary or portfolio, the instructor will still manage the construction of them to some extent; however, their form and content may also be free and unpredictable.

Several studies invoke informal methods, such as anecdotal evidence, informal discussions with students, and informal contacts with alumni. Many studies draw heavily on tutors' reflections, highlighting that the judgement of the instructor will be important in evaluating the efficacy of a course. In addition, several studies invoke indirect measures of success, such as requests for supervisions for independent study projects; the longevity of programme and enrolments, local reputation, and the quality of staff-student relations. Finally, special additional evaluation methods were used, for instance student questionnaires; tracer studies of graduate employability; and outside reviewer reports.

To exemplify the alternative approach, it is worth considering Stilwell's study (2011) of his own pluralist first year course at the University of Sydney. Stilwell's pluralist approach is also evident in the assessment he uses, which asks explicitly students to contrast different approaches. Stilwell's empirical analysis is detailed and extensive. He adopts a mixed-methods approach, utilising several data sources, types and forms of analysis to reach his conclusions. In this respect, at least, Stilwell's paper represents an exemplar of realistic, case-based evaluation. Stilwell demonstrates that the pluralist approach performs well in terms of both liberal and instrumental aims, engaging students, generating employability and achieving changes in critical capacities, flexibility of thought and attitude change.

Critique

It is opportune to evaluate briefly this alternative approach. The critique here is necessarily shorter than that of experiments: we have already considered criticisms of qualitative research; and because experimentalism dominates in economic education research, it was worthy of a longer, more thorough treatment. The evaluation is also lopsided because some of the tests set for experimentalism do not apply to these alternative approaches. For example, it is clear that case-based methods do not usually offer general results; they do not intend that. Similarly, in a mixed-methods approach, there is no default method, so this approach ought not be guilty of that.

It was argued above that experiments may not offer strong explanations, merely statistical generalisations. Do the alternative methods offer better explanations? This question requires empirical evidence. However, *prima facie*, if we take the evidence from case studies as valid, they should offer at least a deeper exploration of specific cases and they ought to help the teacher know *why* whether their intervention 'worked' or did not. Given the concept of relatability, it may be that cases can be connected, creating wider understanding than that from simply one case.

Another criticism of experiments is that they are best suited to short-run assessments of student achievement. This follows because they require that confounding factors are not present, and this requires a short time frame for their data collection. Also, they tend to favour tests at the end of a teaching period, which tend to assess short-run understanding. Again, *prima facie*, one might expect case-based methods to fare better; but again, partly because they do not tend to make claims about the constancy of confounding factors. Learning portfolios and other more qualitative methods may also capture longer-term effects, but of course nothing precludes their use in experimental design.

A final strong critique of experimental design is that it creates ethical issues of (dis)advantaging one group. However, of course this concern applies to any action research study, and so would apply to the alternative methods discussed here. The main difference is that in the latter methods, it is usually the intention to offer the entire cohort the potentially beneficial treatment: there is no *requirement* to create a disadvantaged group.

5. Conclusions

This paper has discussed two broad approaches to evaluation of economics curricula: experimental, and case-based, realistic, mixed-methods approaches. It has been argued that though the experimental approach dominates in economics education, it exhibits notable flaws and may only be suitable in specific contexts. Furthermore, alternative or complementary methods are available. These other methods are rooted in research tradition(s) rather different from those underpinning experiments. Those approaches use a variety of (often qualitative) data types and seek not to create universal or general findings: they respect the uniqueness of the case. Nonetheless, they do seek findings which may translate to other cases. Thus, they have notable advantages in many cases.

This paper argues that educational environments constitute a complex ontology, in which single conceptions of the reality, single theories and, therefore, single methods of evaluation are unlikely to yield conclusive answers. In such environments, mixed-methods methodology may be desirable (Downward and Mearman, 2007). Indeed, Davies and Guest (2010) have argued for various types of evidence in economics education research. There is considerable scope for *triangulation* of data types, sites, theories and analytical methods.

So, to answer the question posed at the outset: economics curricula should be evaluated using a method or set of methods most appropriate to the particular object case. This may involve experiments or elements of the experimental method; but if so, the limitations of that method must be recognised and the conclusions drawn from it, must be utilised with caution. In addition, alternative or complementary methods of evaluation may be usefully deployed.

References

Amin, R. and M. Haneef (2011). 'The quest for better economics graduates: revising the pluralist approach in the case of the International Islamic University, Malaysia', *International Journal of Pluralism and Economics Education*, 2 (1): 96-113.

Barankay, I, M. Johannesson, J. List, R. Friberg, M. Liski and K. Storesletten (2013). 'Guest Editors' preface to the Special Symposium on field experiments', *Scandinavian Journal of Economics*, 115 (1): 1-2.

Barone , C. (2011). 'Contending economic perspectives at a liberal arts college: a 25-year retrospective', *International Journal of Pluralism and Economics Education*, Vol. 2, 1: 19-38.

_____ (2012). 'Student Attitudes Toward Economic Pluralism: Survey-Based Evidence', *International Journal of Pluralism and Economics Education*, forthcoming.

Bassey, M. (2007). 'On he kinds of research in educational settings' in M. Hammersley (ed.) *Educational research and evidence-based practice*, London: Sage.

Bolgar, H. (1965). 'The case study method' in B. Wolman (ed.) *Handbook of Clinical Psychology*, New York: McGraw Hill.

Bonesrønning, H. And L. Opstad (2012). 'How much is students' college performance affected by quantity of study?', *International Review of Economics Education*, 11 (2): 46-63.

Borg, W. and M. Gall (1996). *Educational Research; an introduction*, New York: Longman.

Burawoy, M. (1998). 'The extended case method', *Sociological Theory*, 16 (1): 4-33.

Byrne, D. (2009). 'Complex realist and configurational approaches to cases: a radical synthesis', in Byrne, D. and C. Ragin (eds.) *The Sage handbook of case-based methods*, London: Sage.

Campbell, D. (1969). 'Reforms as experiments', *American Psychologist*, 24: 409-429.

_____ and J. Stanley (1963). Experimental and quasi-experimental evaluations in social research, Chicago: Rand-McNally.

Clarke, P. and A. Mearman (2004). 'Comment on C. Winch, "Economic Aims of Education"', *Journal of Philosophy of Education*, 28 (3).

Contreras, S., F. Badua and M. Adrian (2012). 'Peer effects on undergraduate student performance', *International Review of Economics Education*, 11 (1): 57-66.

Cottrell, S. (2011). *Critical Thinking Skills; Developing Effective Analysis and Argument*, London: Palgrave Macmillan.

Creswell, J. and V. Plano Clark (2011). *Designing and conducting mixed methods research*, Thousand Oaks: Sage.

Cronbach, L. (1982). *Designing evaluations of educational and social programs*, San Francisco: Jossey-Bass.

Davies, P. and R. Guest (2010). 'What effects do we really have on students' understanding and attitudes? How do we know?', *International Review of Economics Education*, 9(1): 6-9.

Denzin, N. (1970). *The research act in sociology: A theoretical introduction to sociological methods*, London: Butterworths.

Downward, P. and Mearman, A. (2007) 'Retroduction as mixed-methods triangulation in economic research: reorienting economics back into social science', *Cambridge Journal of Economics*, 31(2): 77-99.

Earl, P. (2000) 'Indeterminacy in the Economics Classroom', in Earl, P. and S. Frowen, (Eds.): *Economics as an Art of Thought: Essays in Memory of G.L.S. Shackle*, pp. 25-50, London: Routledge.

Elliott, J. (2007). 'Making evidence-based practice educational' in M. Hammersley (ed.) *Educational research and evidence-based practice*, London: Sage.

Gage, N. (2007). 'The paradigm wars and their aftermath' in M. Hammersley (ed.) *Educational research and evidence-based practice*, London: Sage.

Goertz, G. and J. Mahoney (2009). 'Scope in case study research', in Byrne, D. and C. Ragin (eds.) *The Sage handbook of case-based methods*, London: Sage.

Goetz, J. and M. LeCompte (1984). *Ethnography and qualitative design in educational research*, Orlando, FL: Academic Press.

Goode, W. And P. Hatt (1952). *Methods in social research*, New York: McGraw-Hill.

Guba, E. and Y. Lincoln (1982). 'Epistemological and methodological bases of naturalistic inquiry', *Educational Communication and Technology Journal*, 30: 233-52.

Guarasci, R. (2001). 'Developing the Democratic Arts,' *About Campus*, January-February: 9-15.

Hammersley, M. (2007). 'A reply to Hargreaves' in M. Hammersley (ed.) *Educational research and evidence-based practice*, London: Sage.

Hargreaves, D. (1997). 'In defence of evidence-based teaching', *British Educational Research Journal*, 23: 405-419.

_____ (1999). 'Revitalising educational research: lessons from the past and proposals for the future', *Cambridge Journal of Education*, 29 (2): 239-249.

Harvey, D. (2009). 'Complexity and case', in Byrne, D. and C. Ragin (eds.) *The Sage handbook of case-based methods*, London: Sage.

Hodkinson, P., G. Biesta and D. James (2007). 'Learning cultures and a cultural theory of learning', in D. James and G. Biesta (Eds.) *Improving learning cultures in further education*, London: Routledge.

Holland, P. (1986). 'Statistics and causal inference', *Journal of the American Statistical Association*, 81: 945-70.

James, D., G. Biesta, P. Hodkinson, K. Postlethwaite and D. Gleeson (2007) 'Improving learning cultures in Further Education?' in D. James and G. Biesta (Eds.) *Improving learning cultures in further education*, London: Routledge.

Kirby, E., S. Kirby, and M. Lewis (2002). 'A study of the effectiveness of training proactive thinking', *Journal of Applied Social Psychology*, 32 (7): 1538-49.

Lapidus, J. (2011). 'But which theory is right? Economic pluralism, developmental epistemology and uncertainty', *International Journal of Pluralism and Economics Education*, 2 (1): 82-95.

Lawson, T. (1997). *Economics and reality*, London: Routledge.

Lim, C. P. (1998) 'The effect of Computer-Based Learning (CBL) support package on the learning outcome of low-performance economics students', *Computers in Higher Education Economics Review*, 12(1): 19–26.

McIntyre, R. and R. van Horn (2011). 'Contending Perspectives in One Department', *International Journal of Pluralism and Economics Education*, 2 (1): 69-81.

Mahoney, J. and P. L. Terrie (2009). in Byrne, D. and C. Ragin (2009). *The Sage handbook of case-based methods*, London: Sage.

Martinson, R. (1974). 'What works? Questions and answers about prison reform', *Public Interest* 35, 22-45

Mjøset, L. (2009). 'The Contextualist approach to social science methodology', in Byrne, D. and C. Ragin (Eds.) *The Sage handbook of case-based methods*, London: Sage.

Morrison, K. (2009). *Causation in educational research*, London: Routledge.

Oakley, A. (2007). 'Evidence-informed policy and practice: challenges for social science' in M. Hammersley (ed.) *Educational research and evidence-based practice*, London: Sage.

O'Donnell, R. (2009). 'Economic pluralism and skill formation: adding value to students, economies, and societies', in Garnett, R., E. Olsen, and M. Starr (Eds.) *Economic Pluralism*, London, Routledge.
Opie, C. (2004). 'Research approaches' in C. Opie (Ed.) *Doing educational research: a guide for first-time researchers*, London: Sage.
Pawson, R. and N. Tilley (1997). *Realistic Evaluation*, London: Sage.
Postlethwaite, K. (2007). 'Methodological appendix', in D. James and G. Biesta (Eds.) *Improving learning cultures in further education*, London: Routledge.
Punch, K. (2009). *Introduction to Research methods in education*, London: Sage.
Ragin, C. and H. Becker (eds) (1992). *What is a case? Exploring the foundations of social enquiry*, New York: Cambridge University Press.
Resnick, S. and R. Wolff (2011). 'Teaching economics differently by comparing contesting theories', *International Journal of Pluralism and Economics Education*, 2 (1): 57-68.
Rihoux, B. and B. Lobe (2009). 'The case for Qualitative Comparative Analysis', in Byrne, D. and C. Ragin (eds.) *The Sage handbook of case-based methods*, London: Sage.
Schofield, J. (2007). 'Increasing the generalizability of qualitative research' in M. Hammersley (ed.) *Educational research and evidence-based practice*, London: Sage.
Siakantaris, N. (2000). 'Experimental economics under the microscope', *Cambridge Journal of Economics*, 24: 267–281
Sikes, P. (2004). 'Methodology, procedures and ethical concerns' in C. Opie (Ed.) *Doing educational research: a guide for first-time researchers*, London: Sage.
Soh, L. (2001). *A classroom case study on ways to create a motivating learning environment*, unpublished MEd thesis, University of Sheffield.
Stilwell, F. (2011). 'Teaching a pluralist course in economics: the University of Sydney experience', *International Journal of Pluralism and Economics Education*, 2 (1): 39-56.
Tripp, D. (1993). *Critical Incidents in Teaching*, London: Routledge.
Watts, M. and R. Guest (2010). 'Editorial', *International Review of Economics Education*, (special issue on experimental economics in the classroom) in teaching, 9 (2): 6-9.
Zepke, N. (2011). 'Understanding teaching, motivation and external influences in student engagement: how can complexity thinking help?', *Research in Post-Compulsory Education*, 16 (1): 1-13.

Author biography

Andrew Mearman is Associate Professor in Economics and Associate Head of Department (Economics Subject Leader) at the University of the West of England, Bristol. He is an Associate of the Economics Network. He has published widely on economics education and economic methodology, and has focused on mixed or pluralist ways of doing, teaching and evaluating economics.

Contact details

Department of Accounting, Economics and Finance,
UWE, Bristol
Coldharbour Lane
Bristol
BS16 1QY
Email: Andrew.Mearman@uwe.ac.uk

Recent UWE Economics Papers
See http://www1.uwe.ac.uk/bl/bbs/bbsresearch/economics/economicspapers.aspx for a full list

2013

1306 **How should economics curricula be evaluated?**
Andrew Mearman

1305 **Temporary employment and wellbeing: Selection or causal?**
Chris Dawson, Don J Webber and Ben Hopkins

1304 **Trade unions and unpaid overtime in Britain**
Michail Veliziotis

1303 **Why do students study economics?**
Andrew Mearman, Aspasia Papa and Don J. Webber

1302 **Estimating regional input coefficients and multipliers: The use of the FLQ is not a gamble**
Anthony T. Flegg and Timo Tohmo

1301 **Liquidity and credit risks in the UK's financial crisis: How QE changed the relationship**
Woon Wong, Iris Biefang-Frisancho Mariscal, Wanru Yao and Peter Howells

2012

1221 **The impact of the quality of the work environment on employees' intention to quit**
Ray Markey, Katherine Ravenswood and Don J. Webber

1220 **The changing influence of culture on job satisfaction across Europe: 1981-2008**
Gail Pacheco, De Wet van der Westhuizen and Don J. Webber

1219 **Understanding student attendance in Business Schools: an exploratory study**
Andrew Mearman, Don J. Webber, Artjoms Ivļevs, Tanzila Rahman & Gail Pacheco

1218 **What is a manufacturing job?**
Felix Ritchie, Andrew D. Thomas and Richard Welpton

1217 **Rethinking economics: Logical gaps – empirical to the real world**
Stuart Birks

1216 **Rethinking economics: Logical gaps – theory to empirical**
Stuart Birks

1215 **Rethinking economics: Economics as a toolkit**
Stuart Birks

1214 **Rethinking economics: Downs with traction**
Stuart Birks

1213 **Rethinking economics: theory as rhetoric**
Stuart Birks

1212 **An economics angle on the law**
Stuart Birks

1211 **Temporary versus permanent employment: Does health matter?**
Gail Pacheco, Dominic Page and Don J. Webber

1210 **Issues in the measurement of low pay: 2010**
Suzanne Fry and Felix Ritchie

1209 **Output-based disclosure control for regressions**
Felix Ritchie

1208 **Sample selection and bribing behaviour**
Timothy Hinks and Artjoms Ivļevs

1207 **Internet shopping and Internet banking in sequence**
Athanasios G. Patsiotis, Tim Hughes and Don J. Webber

1206 **Mental and physical health: Reconceptualising the relationship with employment propensity**
Gail Pacheco, Dom Page and Don J. Webber

1205 **Using student evaluations to improve individual and department teaching qualities**
Mary R. Hedges and Don J. Webber

1204 **The effects of the 2004 Minority Education Reform on pupils' performance in Latvia**
Artjoms Ivļevs and Roswitha M. King

1203 **Pluralist economics curricula: Do they work and how would we know?**
Andrew Mearman

1202 **Fractionalization and well-being: Evidence from a new South African data set**
Timothy Hinks

1201 **The role of structural change in European regional productivity growth**
Eoin O'Leary and Don J. Webber