Kirkpatrick's levels and education 'evidence'

Sarah Yardley[1] & Tim Dornan[2]

OBJECTIVES This study aims to review, critically, the suitability of Kirkpatrick's levels for appraising interventions in medical education, to review empirical evidence of their application in this context, and to explore alternative ways of appraising research evidence.

METHODS The mixed methods used in this research included a narrative literature review, a critical review of theory and qualitative empirical analysis, conducted within a process of cooperative inquiry.

RESULTS Kirkpatrick's levels, introduced to evaluate training in industry, involve so many implicit assumptions that they are suitable for use only in relatively simple instructional designs, short-term endpoints and beneficiaries other than learners. Such conditions are met by perhaps one-fifth of medical education evidence reviews. Under other conditions, the hierarchical application of the levels as a critical appraisal tool adds little value and leaves reviewers to make global judgements of the trustworthiness of the data.

CONCLUSIONS Far from defining a reference standard critical appraisal tool, this research shows that 'quality' is defined as much by the purpose to which evidence is to be put as by any invariant and objectively measurable quality. Pending further research, we offer a simple way of deciding how to appraise the quality of medical education research.

Medical Education 2012: ***46:*** *97–106*
doi:10.1111/j.1365-2923.2011.04076.x

[1]Keele University Medical School, Faculty of Health, Keele, UK
[2]School of Health Professions Education, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, the Netherlands

Correspondence: Sarah Yardley, Keele University Medical School, Faculty of Health, Keele ST5 5BG, UK. Tel: 00 44 1782 734694; Fax: 00 44 1782 734637; E-mail: syardley@doctors.org.uk

INTRODUCTION

There is a move to make medical education more evidence-based,[1] exemplified by the activities of the Best Evidence Medical Education collaboration (BEME [http://www2.warwick.ac.uk/fac/med/beme]) and other evidence reviews exemplified in Table 1. The BEME collaboration has published 14 reviews to date. Seven of them, listed in Table 1, used Kirkpatrick's levels to appraise evidence, as did the seven other recent non-BEME reviews identified by a literature search, which indicates that these levels are widely used to evaluate education. This paper aims to review their utility as a standard by reviewing their origins, context of use and application in the domain of medical education and beyond.

In a recent book, Donald Kirkpatrick explains how he arrived at the set of four descriptors that are now widely used to evaluate the impact of interventions in education.[2] He had observed that technical training could be evaluated by measuring learners' reactions, learning and behaviour, and their impact on the organisations for which the learners worked.[3] Kirkpatrick's purpose was to provide managers with promptly identifiable and easy-to-measure outcomes in learners and the organisations for which they worked. Business leaders needing tangible evidence that training would enhance their sales volume, product quality and profitability quickly implemented his ideas. Reports of their successful use in business attracted interest from other fields and his ideas spread. Kirkpatrick himself said there was no need to validate the descriptors because accolades poured in.[2] Despite the wide use of Kirkpatrick's levels in medical education, there has been no review or critique of their use in this context. Therefore, we set out to:

1 undertake a narrative review of Kirkpatrick's original writings, subsequent refinements of his work, and publications critiquing the application of his levels;

2 examine how Kirkpatrick's levels have been used in systematic reviews of medical education and examine what is lost by excluding evidence on account of its Kirkpatrick level, and

3 consider alternative approaches to appraising evidence about education.

METHODS

The project was not submitted for research ethics approval because it did not directly involve human subjects or animals. Because our conclusions could have been influenced by our individual experiences of undertaking systematic review and our interpretations of published work, we adopted the principles of cooperative inquiry to help us remain aware of our subjective reactions while working together.[4–6] Epistemologically aligned to constructionism, this methodology entails co-constructing an interpretation by discussing findings, critically reflecting on them cooperatively, and expanding ideas through interactive critique.[7,8]

Study design

We agreed a research focus, research questions, propositions to explore, and initial actions to expand our ideas. In accordance with the principles of cooperative inquiry, we agreed how to carry out actions whilst observing and recording the process and outcome of our experiences. We maintained an audit trail of the developing interpretation by, initially, making notes of face-to-face meetings and using strands of e-mail correspondence as a record. The increasing complexities of the project led us to audio-record face-to-face meetings and, later, use a wiki/web authoring technology to co-construct an interpretation of the evidence we reviewed. Two actions were taken to achieve the first of our three research aims. One was to review Kirkpatrick's original papers and recent reflections on his work.[2] The other was to search for critiques of the application of Kirkpatrick's levels in medical education and – having found no published evidence within the field – beyond. To that end, we searched MEDLINE, CINAHL (Cumulative Index to Nursing and Allied Health Literature), EMBASE, ERIC (Educational Resources Information Centre), BEI (British Education Index), PsycINFO, Academic Search Elite and Business Source Elite for articles with 'Kirkpatrick' in their titles and/or abstracts and which provided a critique of Kirkpatrick's work. Only three were identified, although thousands of articles used (often secondary) references to the levels as a *de facto* standard. We identified Kirkpatrick's own work by searching the same databases extensively and tracing his own references. We chose a narrative review methodology to make maximum use of the little identifiable critique to evaluate the application of Kirkpatrick's levels. We took two actions to achieve the second aim. Having found that the term 'Kirkpatrick' invariably appeared in the abstract of papers using his levels, we searched the journals *Medical Education*, *Medical Teacher*, *Academic Medicine*, *Teaching and Learning in Medicine*, *Advances in Health Sciences Education*, *BMC Medical Education*,

Table 1 Distribution of Kirkpatrick levels in reviews

		Kirkpatrick levels					
		1	2a	2b	3	4a	4b
		Participation	Attitudes	Knowledge and/or skills	Behaviour	Organisational practice	Benefit to patients
Issenberg *et al.*[32]	Simulation education	Kirkpatrick levels used but distribution not shown					
Dornan *et al.*[9]	Early workplace experience (%)	66 (24)	84 (30)	102 (34)	25 (9)	6 (2)	3 (1)
Steinert *et al.*[33]	Faculty development (%)	39 (28)	19 (14)	31 (23)	38 (28)	7 (5)	3 (2)
Hammick *et al.*[34]	Interprofessional education (%)	14 (27)	12 (24)	11 (22)	6 (12)	3 (6)	5 (10)
Driessen *et al.*[35]	Portfolios (%)	19 (90)			2 (10)		
Overeem *et al.*[21]	Formative assessment of doctors' performance (%)	8 (33)	4 (17)		12 (50)		
Colthart *et al.*[36]	Self-assessment (%)	38 (100)					
Tochel *et al.*[37]	Portfolios (%)	7 (16)	26 (58)		10 (22)	2 (4)	
Buckley *et al.*[38]	Portfolios (%)	59 (86)	4 (6)	5 (7)	1 (1)		0
Hill *et al.*[22]	Resident-as-teacher programmes (%)	9 (31)			17 (59)	2 (7)	1 (3); student, not patient
Cherry *et al.*[20]	Asepsis complicating catheter insertion (%)				25 (40)		37 (60)
Wong *et al.*[39]	Internet-based medical education (%)	209 (61)	124 (36)		7 (2)		1 (< 1)
Wong *et al.*[40]	Effect of patient safety and quality improvement education: undergraduate curricula (%)	7 (24)	8 (29)	9 (31)	3 (10)	1 (3)	1 (3)
	Effect of patient safety and quality improvement education: postgraduate curricula (%)	7 (14)	14 (28)	14 (28)	2 (4)	12 (24)	1 (2)
Miller & Archer[41]	Impact of workplace-based assessment on doctors' education and performance (%)	8 (44)	5 (28)	1 (6)	4 (22)*		

* Miller and Archer accepted self-reports of change in behaviour as level 3

British Medical Journal, *Lancet* and the *Journal of the American Medical Association* for review articles published during 2005–2010 that used the word 'Kirkpatrick' in the title or abstract. These are presented in Table 1. The other action was to use a case–control study design to re-examine all publications excluded from our own previous BEME review[9,10] on the grounds of a Kirkpatrick level of ≤ 2. For each excluded publication, we sought two control publications from the same journal in the same year that were included in the final dataset of our previous review. We read all three index publications and their five respective control publications (there were five rather than six control publications because only one control was available for one of the three index publications) and discussed whether it was self-evident that they should have been excluded or included, and in what ways they illuminated the review topic. This identified deficiencies in Kirkpatrick's levels that, together with the previous steps, helped us achieve the third aim of considering alternatives to Kirkpatrick's levels.

RESULTS

The suitability of Kirkpatrick's levels for appraising education interventions

Most articles found by our search used Kirkpatrick's levels as heuristics in education evaluation; just four critiqued their use[11–14] and one of these found that Kirkpatrick's levels were applied uncritically in the field of human resource development.[14] Abernathy,[12] noting that the levels could influence the questions asked and results produced, rejected them as unsuitable for evaluating either 'soft' outcomes or continuous learning (as opposed to time-limited interventions). Alliger and Janak identified three types of assumption by which Kirkpatrick's model could tacitly shape research findings, comprising: assumptions of hierarchy associated with the numeric labelling of levels; assumptions of causal links between levels, and assumptions that the levels are positively inter-correlated.[11] Blanchard *et al.*[13] argued that the purpose of any research had to be determined before any evaluation of it at any particular Kirkpatrick level could be considered.

Although none of those studies concerned medical education, they seem applicable to it and Kirkpatrick himself might have agreed with these authors because he actually advocated using the levels as a training heuristic,[2] not to evaluate how professionals become expert practitioners through deliberate practice and social learning. He chose the levels to measure very short-term and tangible endpoints like sales volume, quality and profitability. Kirkpatrick's solution to intangible benefits of training, which he acknowledged in his original work, was to link them to tangible benefits because training orientated towards specific measurable behaviours could be assigned a market value.[2] Of his numerous references to successful applications of the levels,[2] none came from a field as complex as medical education, which differs from business in that it is required to meet the needs, equitably, of a whole array of beneficiaries, including patients, students, practitioners, communities and health care organisations. A problem with Kirkpatrick's levels is that different levels concern different beneficiaries: levels 1–3 concern learners; level 4a concerns organisations, and level 4b concerns patients. Teachers are missing from the scheme altogether. The model does not allow for the rich variety of outcomes that can be evaluated using qualitative as well as quantitative methodologies, nor explain how or why such outcomes are consequential to particular elements of complex interventions. It tends only to be used to measure anticipated outcomes and ignores unanticipated consequences. That is, it asks 'Was outcome X achieved as intended, or not?' rather than 'What were the outcomes of this intervention?' A clinical parallel would be a clinical trial that measured only the intended effects of a new drug and not its side-effects.

Application of Kirkpatrick's levels to medical education research

Opinion was expressed in the late 1990s that medical education was not using evidence in a way that would most effectively support practice and, as we have described elsewhere, the BEME collaboration came into being.[15] Its mission was to conduct 'a logical, explicit, and comprehensive appraisal of available information to determine the best evidence relating to an issue in medical education'.[16] Forty years after Kirkpatrick's original work, the BEME collaboration adopted a modified version of Kirkpatrick's levels (which it named a 'hierarchy') as a grading standard for bibliographic reviews (Table 2). A prototype coding sheet, accompanied by explanatory notes, offered two complementary ways of appraising evidence, using either Kirkpatrick's 'hierarchy' to grade the impact of interventions (Table 2) or a simple anchored rating scale of 1–5 of the 'strength' (Table 3) or trustworthiness of findings. The BEME's use of the term 'hierarchy' implied that a higher Kirkpatrick level represented greater quality.

Table 2 Kirkpatrick's levels as represented on the Best Evidence Medical Education Collaboration's specimen coding sheet (http://www2.warwick.ac.uk/fac/med/beme/writing/resources/appendix_iiia_beme_coding_sheet.pdf)

Impact of intervention studied

Code the level of impact being studied in the item and summarise any results of the intervention at the appropriate level. Note: include both predetermined and unintended outcomes

Kirkpatrick hierarchy

- Level 1 Participation: covers learners' views on the learning experience, its organisation, presentation, content, teaching methods, and aspects of the instructional organisation, materials, quality of instruction
- Level 2a Modification of attitudes/perceptions: outcomes relate to changes in the reciprocal attitudes or perceptions between participant groups towards the intervention/simulation
- Level 2b Modification of knowledge/skills: for *knowledge*, this relates to the acquisition of concepts, procedures and principles; for *skills* this relates to the acquisition of thinking/problem-solving, psychomotor and social skills
- Level 3 Behavioural change: documents the transfer of learning to the workplace or willingness of learners to apply new knowledge and skills
- Level 4a Change in organisational practice: wider changes in the organisation or delivery of care, attributable to an educational programme
- Level 4b Benefits to patient/clients: any improvement in the health and well-being of patients/clients as a direct result of an educational programme

Table 3 Appraisal of the strength of medical education research (http://www2.warwick.ac.uk/fac/med/beme/writing/resources/appendix_iiia_beme_coding_sheet.pdf)

Strength

1 No clear conclusions can be drawn; not significant
2 Results ambiguous, but there appears to be a trend
3 Conclusions can probably be based on the results
4 Results are clear and very likely to be true
5 Results are unequivocal

The BEME scheme has been widely adopted, largely because the notion of a hierarchy of evidence resonates strongly with two dominant themes in clinical medicine, the parent discipline of medical education, namely: the 'evidence hierarchy' (from the case study at the bottom to the statistical meta-analysis at the top) of evidence-based medicine,[17] and the valuing of 'hard' clinical outcomes over 'soft' intermediate outcomes in contemporary health services research, against which medical education research has been unfavourably compared.[18,19]

Our own first BEME review[9] (of early workplace experience in undergraduate medical education) used the levels and, accepting them as a hierarchy, treated a higher Kirkpatrick level as indicative of a more important outcome. This review found that 24% of outcomes were at level 1, which we then regarded as unimportant, and the other 76% were progressively more important according to their higher levels. A total of 64% of outcomes were found to be at level 2, leaving only 12% at levels 3 and 4 combined. When we added in an appraisal of the 'strength' of outcomes (Table 3), only 42% of published outcomes were both strong (rated at ≥ 3) and important (Kirkpatrick level ≥ 2) and then mostly at level 2 in the hierarchy. Early workplace experience in undergraduate medical education, those descriptors seemed to tell us, was supported by 'little good evidence'.[9] A recent update of the review found a fall of 50% per annum in the number of articles published, but such a marked increase in the proportion of 'strong and important' outcomes that the production of new 'best evidence' had fallen relatively little. Either a lower volume of higher-quality research was being undertaken or researcher interests and editorial standards had become more stringently orientated to such outcomes.[10]

Table 1 shows the comprehensive set of articles that used Kirkpatrick's levels in medical education evidence synthesis. It includes six of 14 BEME reviews. It shows we were not alone in finding relatively few Kirkpatrick level 3 or 4 outcomes. In only three of 14 data analyses (21%) were half or more of the outcomes rated at a level > 2. In one of them, an investigation by Cherry *et al.*[20] into the impact of educational interventions on aseptic insertion and the maintenance of central venous catheters in acute care, over half the outcomes were rated at level 4. In two others, a review of the outcomes of formative assessment of doctors' performance[21] and a review of resident-as-teacher programmes,[22] 50% or more of the outcomes were rated at level 3. Careless catheter insertion can cause patients to develop septicaemia and its complications within hours, and training in this practical procedure can reasonably be expected to bring about an immediate reduction in life-threatening infections. Similarly, formative assessment of doctors' behaviour and teaching

residents to teach might reasonably be expected to bring about immediate behavioural changes. Thus, level 3 or 4 outcomes were reasonable expectations in these particular reviews, given the nature of the interventions and their anticipated outcomes. At the other extreme, early workplace experience, for example, might take months or even years to have any demonstrable effect on learners, let alone patients, and its main effects might be on learners' attitudes, the benefits of which to patients would be very tricky to measure. Attempts to measure comparable effects of early experience on patients would just not make sense.

Most papers in Table 1 described what learners experienced (level 1) or measured what they learned (levels 2a and 2b); these have been more simply termed 'description' and 'justification' studies, respectively, and each has its own value.[23] The snag is that outcomes in 'clarification studies', which are a rich basis on which to strengthen medical education,[23] could fit under any or all of Kirkpatrick's levels. Yet, unless we understand how, and why, effects are consequential to particular elements or interactions, it will be difficult to refine education to maximise benefit. To give a specific example, it is possible that a study clarifying how an educational intervention affected learners' emotions could be classified as demonstrating outcomes at level 1 (reactions) or 2 (attitudes), which are regarded as relatively unimportant, despite being self-evidently important to the professional development of the learners. Are outcomes necessarily more important than processes (which are not included in Kirkpatrick's levels)? How can the extent to which empirical research is theoretically grounded, or even the value of purely theoretical scholarship, be given due recognition by a classification restricted to outcomes? We are not alone in our criticism of the use of Kirkpatrick's levels to stratify evidence. Holton criticised their use as a hierarchy on the grounds that they lack important attributes of a theory and lack supportive evidence to indicate that lower-level outcomes are prerequisite to higher-level ones.[14]

Our second empirical investigation of Kirkpatrick's levels in medical education research found that all three of the papers excluded from our early workplace experience review because the Kirkpatrick level of their outcomes was 1 (learners' reactions)[10] contained information that could answer valid review questions, albeit questions that differed from that we sought to answer at the time. This information was relevant for policymakers,[24] curriculum designers,[25] and those interested in the development of medical students as researchers[24] or in how students use narrative to make sense of their experiences.[26]

Alternatives for appraising research in medical education

The three (21%) BEME systematic reviews that had a substantial proportion of higher-level ratings show there is a role for Kirkpatrick's levels, albeit a limited one. When evaluating relatively simple training interventions, the outcomes of which emerge rapidly and are easily observed within classical experiment designs, the levels can direct attention to important beneficiaries other than learners (notably patients). The preceding review, however, leads us to conclude they are unsuitable for the higher proportion of education interventions, which are complex, in which the most important outcomes are longer-term, and in which process evaluation is as important as (perhaps even more important than) outcome evaluation. Indeed, our review found a body of opinion that considered that Kirkpatrick's levels, applied to the wrong type of evidence, might be harmful.[11–14] What alternative ways are there, then, to critique the quality of various types of evidence in a scholarly way without allowing the type of evidence to bias its evaluation? Put another way, how do we balance the right level of inclusiveness with rigour in our approach to value? It is important that the current state of knowledge, including 'negative' findings and specific needs for new or more rigorous work to usefully inform further research or practice innovation, is represented.

The scholarship of systematic review in clinical science takes its origins from a paper published 40 years ago by the epidemiologist Archie Cochrane, in which he berated medical practice for being ineffective or frankly harmful.[27] The Cochrane Collaboration (http://www.cochrane.org) came into existence to promote clinical trials, using systematic review and statistical meta-analysis to synthesise findings from their aggregated results. 'Evidence' was rated as 'weak' or 'strong' according to standard criteria, which appraised its ability to support the statistical estimation of effect sizes. The Cochrane approach is not the only one in the health domain. The Joanna Briggs Institute (http://www.joannabriggs.edu.au/about/home.php) and the W K Kellogg Foundation (http://www.wkkf.org), both of which seek to improve health care practice through multidisciplinary working, have taken a pluralistic approach and do not place randomised controlled trials at the top of a hierarchy, regardless

of the question posed. Recognising that the hypothetico-deductive, experimental approach of natural sciences is 'ill-equipped to help us understand complex, comprehensive, and collaborative community initiatives' (http://www.wkkf.org/knowledge-center/resources/2010/w-k-kellogg-Foundation-Evaluation-Handbook.aspx), they allow questions to be asked and answered without forcing complex systems to fit the evaluative tools of one dominant research paradigm. By contrast, the Campbell Collaboration (http://www.campbellcollaboration.org), which reviews evidence related to education, crime and justice, and social welfare, has aligned itself with the Cochrane Collaboration in holding data that are suitable for statistical meta-analysis as of intrinsically higher quality.

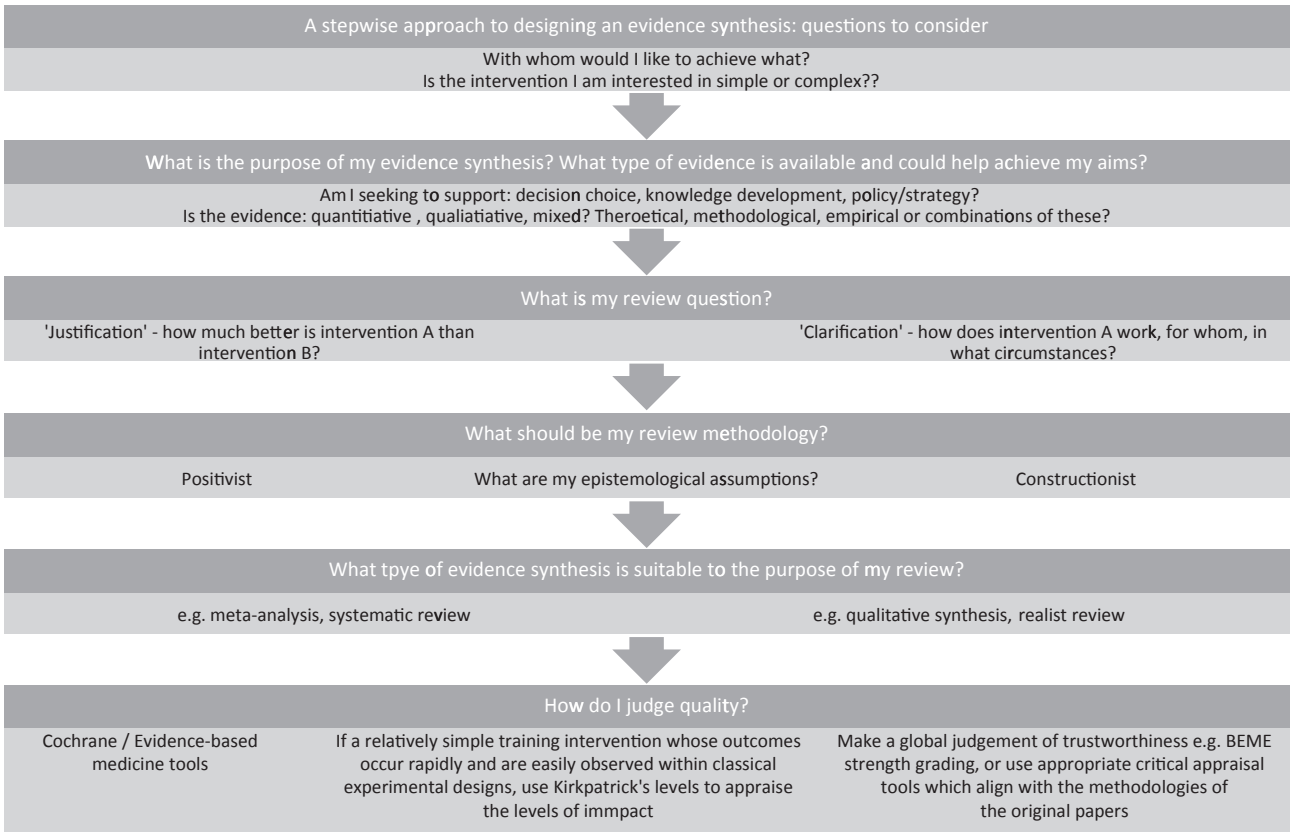
Thus, different review methodologies start from different 'epistemological' assumptions, where the term 'epistemological' refers to the relationship between the knower and the known. The Cochrane approach, drawn from classical scientific methodology, has a positivist epistemology which allows it to reduce complex situations to a comparison of variables within relatively simple experiment designs. Its standards of critical appraisal are consistent with its epistemological stance. Pope *et al.* noted that systematic review, although it is strongly favoured in the clinical domain because it helps in making choices between alternative treatments, is not the only way of synthesising evidence.[28] The Cochrane Collaboration's use of evidence for 'decision support' can be distinguished from the (non-dichotomous) use of evidence for 'knowledge support'. Aggregative or interpretive methods of evidence synthesis that mix qualitative with quantitative evidence, or synthesise qualitative evidence alone, give better knowledge support and start from constructionist rather than positivist epistemological assumptions.[28] Medical education research, our reviews have shown, is pluralistic. So where does that leave the four out of five reviewers whose bibliographic research does not lend itself to Kirkpatrick rating?

Far from defining a reference standard for critical appraisal, this review casts doubt on whether such a standard could ever exist and shows how many questions must be answered when planning an evidence synthesis. Rather than leave the reader with no basis on which to appraise evidence, we conducted a thought experiment in order to define a logical approach. For experimental research conducted on positivist principles, the critical appraisal tools of evidence-based medicine can be applied to education evidence. Under the conditions defined in the first paragraph of this section, such as in the evaluation of relatively simple training interventions, Kirkpatrick's levels are appropriate. In the majority of cases (perhaps 80% of medical education evidence syntheses), a constructionist epistemology is likely to be appropriate, in which case critical appraisal will rest on simple global judgements of trustworthiness, such as the BEME scale of 1–5. Although critical appraisal tools appropriate to individual methodologies could be applied to individual studies included within a review, any gain in reliability is likely to make little difference to the overall conclusions pieced together from multiple different methodologies.

DISCUSSION

The art of evidence synthesis, we conclude, lies in making well-considered choices rather than valorising one methodology or appraisal standard over another, echoing Eva's view that there can be no single arbiter of quality because it is the use to which evidence is put that determines its utility.[29] The use of evidence to support policy, define outcomes, identify new research questions, answer practical teaching questions, inform teachers' personal development, serve as a debating tool or establish the 'state of knowledge' on a subject can all dictate different methodologies. Even the last of these, which is often presented as a neutral assessment, involves ontological and epistemological positioning. If the topic in question is the efficacy of a simple intervention compared with a placebo administered under controlled conditions, a 'naive realist' ontology and epistemology[30] would direct the use of Cochrane critical appraisal standards and estimation of effect sizes. The more reductionist a review, the clearer its results, but perhaps also the less applicable they are. Where medical education really deviates from evidence-based medicine is in its recognition of a wide gap between the results of simple experiments and their applicability in 'real practice'. Context as well as process impacts on educational outcomes. Moreover, rich nuances or even the whole essence of information may be lost when stories of experience are omitted.

For all of these reasons, it is likely a reviewer will need to consider qualitative as well as quantitative sources of evidence and 'construct' an argument fitted to the conversation he or she wants to be part of in the relativist, social world of education practice. If the reviewer wants to influence policy, a



Example A

Researchers who want to review the literature to establish the effectiveness of educational interventions with a clearly defined, easily identifiable and measurable endpoint (e.g. Cherry *et al.*[20]) might address the issues as follows:

'I want to ensure educators use interventions most likely to prevent infection of central venous catheters. It is a relatively simple intervention. Because I hope to influence policy, I will limit my search to direct and structured interventions, which are closely linked to implementation in practice. I will consider any type of evidence but quantitative data are most likely to justify the choosing of one approach over another as 'best practice'. My epistemological assumptions are more positivist than constructionist and so I am aiming to conduct a systematic review (even a meta-analysis if suitable data are available) and I will judge quality on Kirkpatrick's levels and determine the impact of interventions on infection rates on evidence-based medicine principles'.

Example B

Researchers who want to understand how authentic early experience can contribute to the complex educational processes to which a medical student is exposed (e.g. Yardley *et al.*[10]) might, conversely, decide early on that the main aim of their review is to clarify what is known about those processes from different perspectives and data sources in order to identify areas that require further research. They might frame a clarification question and conduct a qualitative synthesis (perhaps converting quantitative into qualitative data), which calls on reviewers to make a judgement of the trustworthiness of data and to integrate theory-based publications with empirical ones. In that case, the synthesis will be the richer for *not* rejecting evidence on the basis of Kirkpatrick's levels. The whole approach is justified by adopting constructionist epistemological assumptions.

Figure 1 Questions to consider when designing an evidence synthesis

realist stance and attendant methods may be appropriate,[31] whereby the reviewer uses pragmatic judgement to answer questions like: If I were reading the original papers as a practitioner, what would I take away from them? What would I accept within context or pass judgement on in a more refined or nuanced manner than the current systematic review process allows? How can I stratify the studies on this topic to see where evidence is strongest or limited without unnecessarily discounting partially helpful informa-

tion? Reviewers who seek to position new research may need to seek out previously unsuccessful studies or negative results as well as successful methods and desired results.

Our broad conclusion is that the purpose to which evidence is put influences its trustworthiness and the best way of synthesising it. Having rejected the methodological assumptions of scientific experimentation and the clinical assumption of patient benefit as reference standards of evidence, we suggest that researchers synthesising evidence should: state very clearly the aims of their work; make their epistemological and ontological assumptions explicit; admit any evidence that is appropriate to the aim, including complex and qualitative evidence; consider features of empirical research such as the strength of its theoretical orientation and its relevance to the review question when considering its weight in the final synthesis, and make absolutely transparent, when reporting a review, the decisions they took and their reasons for taking them. Figure 1 outlines approaches reviewers might take pending clearer results from bibliographic research.

Contributors: SY contributed to the conception and design of this study, researched appropriate methods, conducted the subsequent acquisition, analysis and interpretation of data, and fully cooperated in the cooperative inquiry methods employed. TD contributed to the conception and design of the research, identified cooperative inquiry as a potential method and fully cooperated in the method. Both authors contributed to the drafting and revision of this paper and approved the final manuscript for submission.
Acknowledgements: none.
Funding: none.
Conflicts of interest: both authors have previously participated in Best Evidence Medical Education review groups.
Ethical approval: not applicable.

REFERENCES

1 Jason H. The importance – and limits – of best evidence medical education. *Educ Health* 2000;**13**:9–14.
2 Kirkpatrick J, Kirkpatrick W. *Kirkpatrick, Then and Now: A Strong Foundation for the Future*. St Louis, MO: Kirkpatrick Partners 2009.
3 Kirkpatrick DL. Evaluation of training. In: Craig R, Bittel L, eds. *Training and Development Handbook*. New York, NY: McGraw-Hill 1967;87–112.
4 Heron J, Reason P. The practice of co-operative enquiry: research with rather than on people. In: Reason P, Bradbury H, eds. *Handbook of Action Research: the Concise Paperback Edition*. London: Sage 2006;144–54.
5 Heron J. *Co-operative Inquiry: Research into the Human Condition*. London: Sage Publications 1996;1–125.
6 Oates BJ. Cooperative enquiry: reflections on practice. *Electron J Bus Res Meth* 2002;**1**:27–37.
7 Reason P. The practice of cooperative inquiry. *Syst Pract Action Res* 2002;**15**:169–76.
8 Holt NL. Representation, legitimation, and autoethnography: an autoethnographic writing story. *Intern J Qualitative Meth* 2003;**2** (1):18–28.
9 Dornan T, Littlewood S, Margolis S, Scherpbier A, Spencer J, Ypinazar V. How can experience in clinical and community settings contribute to early medical education? A BEME systematic review. *Med Teach* 2006;**28**:3–18.
10 Yardley S, Littlewood S, Margolis SA, Scherpbier A, Spencer J, Ypinazar V, Dornan T. What has changed in the evidence of early experience? Update of a BEME systematic review. *Med Teach* 2010;**32**:740–6.
11 Alliger GM, Janak EA. Kirkpatrick's levels of training criteria: thirty years later. *Pers Psychol* 1989;**42**:331–42.
12 Abernathy DJ. Thinking outside the evaluation box. *Train Dev* 1999;**53**:18–23.
13 Blanchard PN, Thacker JW, Way SA. Training evaluation: perspectives and evidence from Canada. *Int J Train Dev* 2000;**4**:295–304.
14 Holton EF. The flawed four-level evaluation model. *Hum Res Develop Quart* 1996;**7**:5–21.
15 Dornan T, Peile E, Spencer J. On 'evidence'. *Med Educ* 2008;**42**:232–3.
16 Best Evidence Medical Education (BEME) Steering Group. Guide for topic review groups on carrying out BEME systematic reviews. 2003; http://www2.warwick.ac.uk/fac/med/beme/writing/resources/guide_for_prospective_review_groups_pdf. [Accessed 21 October 2010.]
17 Wikipedia. Hierarchy of evidence. http://en.wikipedia.org/wiki/Hierarchy_of_evidence. [Accessed 21 October 2010.]
18 Todres M, Stephenson A, Jones R. Medical education research remains the poor relation. *BMJ*2007;**335**:333–5.
19 Stephenson A, Todres M, Jones R. Reply to Dornan *et al.*'s 'On evidence'. *Med Educ* 2009;**43**:390–1.
20 Cherry MG, Brown JM, Neal T, Shaw NB. What features of educational interventions lead to competence in aseptic insertion and maintenance of CV catheters in acute care? BEME Guide No. 15. *Med Teach* 2010;**32**:198–218.
21 Overeem K, Faber MJ, Arah OA, Elwyn G, Lombarts KMJMH, Wollersheim HC, Grol RP. Doctor performance assessment in daily practice: does it help doctors or not? A systematic review. *Med Educ* 2007;**41**:1039–49.
22 Hill AG, Uu CBM, Hattie J. A systematic review of resident-as-teacher programmes. *Med Educ* 2009;**43**:1129–40.

23 Cook DA, Bordage G, Schmidt HG. Description, justification and clarification: a framework for classifying the purposes of research in medical education. *Med Educ* 2008;**42**:128–33.
24 Zorzi A, Rourke J, Kennard M, Peterson M, Miller K. Combined research and clinical learning make rural summer studentship programme a successful model. *Education for Health* 2005;**18**:329–37.
25 Vieira JE, Nunes MdPT, Martins MdA. Directing student response to early patient contact by questionnaire. *Med Educ* 2003;**37**:119–25.
26 Gutierrez MC, Soto RG. Alligator attack: an illustration of the impact of early clinical exposure. *Med Educ* 2002;**36**:1182–4.
27 Cochrane A. *Effectiveness and Efficiency: Random Reflections on Health Services*. London: Nuffield Provincial Hospitals NHS Trust 1972;1–92.
28 Pope C, Mays N, Popay J. *Synthesising Qualitative and Quantitative Health Evidence*. Maidenhead: Open University Press 2007;1–224.
29 Eva K. Broadening the debate about quality in medical education research. *Med Educ* 2009;**43**:294–6.
30 Guba EG, Lincoln YS. Paradigmatic controversies, contradictions, and emerging confluences. In: Denzin NK, Lincoln YS, eds. *The Sage Handbook of Qualitative Research*, 3rd edn. Thousand Oaks, CA: Sage Publications 2005;191–215.
31 Pawson R. *Evidence-Based Policy. A Realist Perspective*. London: Sage Publications 2006;1–96.
32 Issenberg SB, McGaghie WC, Petrusa ER, Gordon DL, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach* 2005;**27**:10–28.
33 Steinert Y, Mann K, Centeno A, Dolmans D, Spencer J, Gelula M, Prideaux D. A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME Guide No. 8. *Med Teach* 2006;**28**:497–526.
34 Hammick M, Freeth D, Koppel I, Reeves S, Barr H. A best evidence systematic review of interprofessional education: BEME Guide No. 9. *Med Teach* 2007;**29**:735–51.
35 Driessen E, van Tartwijk J, van der Vleuten C, Wass V. Portfolios in medical education: why do they meet with mixed success? *Med Educ* 2007;**41**:1224–33.
36 Colthart I, Bagnall G, Evans A, Allbutt H, Haig A, Illing J, McKinstry B. The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice. BEME Guide No. 10. *Med Teach* 2008;**30**:124–45.
37 Tochel C, Haig A, Hesketh A, Cadzow A, Beggs K, Colthart I, Peacock H. The effectiveness of portfolios for postgraduate assessment and education: BEME Guide No. 12. *Med Teach* 2009;**31**:299–318.
38 Buckley S, Coleman J, Davison I, *et al*. The education effects of portfolios on undergraduate student learning: a Best Evidence Medical Education (BEME) systematic review. BEME Guide No. 11. *Med Teach* 2009;**31**:282–98.
39 Wong G, Greenhalgh T, Pawson R. Internet-based medical education: a realist review of what works, for whom, and in what circumstances. *BMC Med Educ* 2010;**10** (12).
40 Wong B, Etchells E, Kuper A, Levinson W, Shojania KG. Teaching quality improvement and patient safety to trainees: a systematic review. *Acad Med* 2010;**85**:1425–39.
41 Miller A, Archer J. Impact of workplace-based assessment on doctors' education and performance: a systematic review. *BMJ* 2010;**341**:c5064.

Received 22 October 2010; editorial comments to authors 14 December 2010, 26 April 2011; accepted for publication 1 June 2011