



Analysing synthesis of evidence in a systematic review in health professions education: observations on struggling beyond Kirkpatrick

Gillian Maudsley & David Taylor

To cite this article: Gillian Maudsley & David Taylor (2020) Analysing synthesis of evidence in a systematic review in health professions education: observations on struggling beyond Kirkpatrick, Medical Education Online, 25:1, 1731278, DOI: [10.1080/10872981.2020.1731278](https://doi.org/10.1080/10872981.2020.1731278)

To link to this article: <https://doi.org/10.1080/10872981.2020.1731278>



© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 31 Mar 2020.



Submit your article to this journal [↗](#)



Article views: 139




View related articles [↗](#)



View Crossmark data [↗](#)

Analysing synthesis of evidence in a systematic review in health professions education: observations on struggling beyond Kirkpatrick

Gillian Maudsley^a and David Taylor ^{a,b}

^aDepartment of Public Health & Policy, The University of Liverpool, Liverpool, UK; ^bMedical Education & Physiology, College of Medicine, Gulf Medical University, Ajman, United Arab Emirates

ABSTRACT

Background: Systematic reviews in health professions education may well under-report struggles to synthesize disparate evidence that defies standard quantitative approaches. This paper reports further process analysis in a previously reported systematic review about mobile devices on clinical placements.

Objective: For a troublesome systematic review: (1) Analyse further the distribution and reliability of classifying the evidence to Maxwell quality dimensions (beyond 'Does it work?') and their overlap with Kirkpatrick K-levels. (2) Analyse how the abstracts represented those dimensions of the evidence-base. (3) Reflect on difficulties in synthesis and merits of Maxwell dimensions.

Design: Following integrative synthesis of 45 K2–K4 primary studies (by combined content-thematic analysis in the pragmatism paradigm): (1) Hierarchical cluster analysis explored overlap between Maxwell dimensions and K-levels. Independent and consensus-coding to Maxwell dimensions compared (using: percentages; kappa; McNemar hypothesis-testing) pre- vs post-discussion and (2) article abstract vs main body. (3) Narrative summary captured process difficulties and merits.

Results: (1) The largest cluster (five-cluster dendrogram) was acceptability–accessibility–K1–appropriateness–K3, with K1 and K4 widely separated. For article main bodies, independent coding agreed most for appropriateness (good; adjusted kappa = 0.78). Evidence increased significantly pre–post-discussion about acceptability ($p = 0.008$; 31/45→39/45), accessibility, and equity-ethics-professionalism. (2) Abstracts suggested efficiency significantly less than main bodies evidenced: 31.1% vs 44.4%, $p = 0.031$. (3) Challenges and merits emerged for before, during, and after the review.

Conclusions: There should be more systematic reporting of process analysis about difficulties synthesizing suboptimal evidence-bases. In this example, Maxwell dimensions were a useful framework beyond K-levels for classifying and synthesizing the evidence-base.

ARTICLE HISTORY

Received 13 December 2019
Revised 3 February 2020
Accepted 7 February 2020

KEYWORDS

Best evidence; cluster analysis; epistemology; evidence-based education; evidence synthesis; Kirkpatrick levels; Maxwell dimensions of quality; medical education; process analysis; systematic review

Introduction

Synthesizing messy evidence in health professions education can be more art than science, more pragmatism than finesse, and more trouble apparently than it's worth. Nevertheless, exploring the evidence enough to make useful recommendations and to critique and improve the process must be better than investigatory inaction waiting for the perfect evidence-base.

Whether viewed as what is probably, plausibly, or often generally true, the nature of the 'evidence' [1] and the nature of its 'good quality' are debatable concepts. Approaches to synthesis are eclectic in Best Evidence Medical Education (BEME) systematic reviews [2]. In-depth consideration is limited though [3] about how to classify and analyse diverse evidence that thwarts standard quantitative approaches to the systematic review, and qualitative research synthesis is contested [4,5]. In making sense of a difficult synthesis, a post hoc standard veneer of slick reporting might hide how

researchers cycle through confusion and clarity. While Kirkpatrick's four-level outcomes-based model or taxonomy has been popular for imposing some order in systematic reviews in health professions education, it is useful to analyse how other frameworks might enhance the processes of organizing and synthesis.

The popularity of the Kirkpatrick model

The Kirkpatrick model has been popular for straightforward, practical evaluation of training interventions by $K1 = \text{reaction}$, $K2 = \text{learning}$, $K3 = \text{behaviour}$, and $K4 = \text{results}$ [6], especially in medical education [7,8]. Many BEME systematic reviews have used the model to filter and summarize [1], and its use to organize and synthesize systematic reviews and 'quality-score' quantitative evidence extends across diverse health-care education [9–11]. The common inferences that these 'levels' are a causal sequence ($K1 \rightarrow K2 \rightarrow K3 \rightarrow K4$) or

hierarchical in value (K4>K3>K2>K1) have attracted criticism though [7,8].

Whether or not Kirkpatrick meant a causal hierarchy, he did view K1→K4 as increasing in complexity and meaning – thorough evaluation might use all levels [6]. Moreau [12] noted that the New World Kirkpatrick Model [13] improved on three criticisms: – *Difficulties evaluating K3/K4*: Evaluation now also focused on ways of promoting the application of learning (K3) and contributions to organizational goals (K4). – *Ignoring confounding and intervening variables*: Each K-level now included various personal and organizational influences. – *Inferring an unproven causal chain (K1→K2→K3→K4)*: Considering ‘chains of evidence’ rather than implying a causal hierarchy of levels allowed non-sequential use of the levels. Such advances complemented re-thinking evidence synthesis beyond ‘effectiveness = Does it work?’.

Building beyond Kirkpatrick and ‘what works?’, using an example

The ‘What works?’ question could be dismissed as a narrow interpretation of effectiveness that uses only quantitative evidence for justification, or it could be interpreted more widely from whatever evidence supports ‘whether (it works)’ plus: ‘how, why, and in what circumstances (the context)?’. ‘The difficulty of evaluating any educational philosophy in a scientific manner’ occupied the early days of seeking best evidence (available) in medical (and other health professions) education [14, p.1]. The push to widen the horizons of BEME reviews has continued [15]. Systematic reviews in health services research and public health have long since been dealing with condensing complexity into concise counsel.

Petticrew [16, p.2] argued that the question for complex interventions should be:

“What has happened previously when this intervention [has] been implemented across a range of contexts, populations and subpopulations, and how have those effects come about?”

‘Does it work?’ becomes ‘meaningless and usually unanswerable’ (p.2) for complex interventions, and narrative reviews seek evidence to reduce uncertainty rather than to derive a precise effect-size. Even weak studies provide illumination when a field is still in development and:

“evidence synthesis often is, and should be, an exercise in Bayesian decisionmaking, and reducing uncertainty, and not hypothesis testing” (p.5).

BEME review 52 [17] investigated: ‘What works best for health professions students using mobile (hand-held) devices for educational support on clinical placements?’ in an underdeveloped evidence-base. This was about a complex intervention and required versatile

interpretation and classification of a mash-up of ‘whether’ (justification), ‘how/why’ (clarification), and ‘what’ (description) evidence [2,18]. Of the K2–K4 primary empirical studies included (K1-only studies were excluded), 46.7% (21/45) were mixed methods, 33.3% quantitative, and 20.0% qualitative research. K3 (86.7%) and S3-strength evidence (*Conclusions can probably be based on the results*) (55.6%) [19,20] predominated. There were only five L6 (randomized controlled trials) and two L5 (longitudinal) designs [21]. About three-quarters had supplementary K1 evidence and 53.3% had K4 evidence, mostly K4b. Inter-observer agreement on filtering abstracts was good (e.g. 92.1% in the final 2016 update, kappa = 0.64, p < 0.0001).

BEME review 52 concluded about mobile devices as educational support that [17]:

- They supported students’ learning on clinical placement via: assessment; communication; clinical decision-making; logbook or notetaking; and most often accessing information.
- In the hidden and informal curricula, ‘what happened’ was that students were:
 - *bothered about*: actual and perceived disapproval of peers, clinicians or educators, and patients; confidentiality and privacy; and security aspects,
 - *side-tracked by*: social connectivity (or other private use) and hectic clinical settings,
 - *confused by*: policy ambiguity.

That review moved beyond ‘Does it work?’ and beyond K-levels. Much of the synthesis of the required ‘What works best ...?’ question did implicitly answer ‘What has happened previously with use of mobile devices on clinical placements?’. Maxwell’s dimensions from health services research [22,23] then helped to widen horizons. Used in evaluating quality of care, this 3As & 3Es framework considers:

Acceptability (What do users prefer and how satisfied are they?). *Accessibility* (How reachable is the service? What are the barriers?). *Appropriateness* (How relevant is the service to needs?). *Effectiveness* (Does it work? What are the outcomes?). *Efficiency* (How are outputs to inputs? What are costs?). *Equity* (How fair is it?).

Adapting these dimensions (Table 1) helped to organize evidence and deliberate about synthesis. Previous BEME reviews did not feature this framework, warranting further analysis.

In a preliminary analysis, BEME review 52 reported that the commonest Maxwell evidence-profiles (just under one-half) supported accessibility, appropriateness, acceptability, and effectiveness, or those plus efficiency, with little about equity-ethics-professionalism. Further analysis would gainfully explore the usefulness of that additional framework beyond Kirkpatrick, how both frameworks

overlapped, and insights about presenting that body of evidence. As Regehr [24, p.34] argued:

“We are bound to learn more from our own work and that of others if we systematically examine and document our struggles than if we loudly proclaim our successes.”

In extracting transferable messages about the process of a systematic review, the aims here were to (1) Analyse further the distribution and reliability of classifying the evidence to Maxwell quality dimensions (beyond ‘Does it work?’) and their overlap with K-levels. (2) Analyse how the abstracts represented those dimensions of the evidence-base. (3) Reflect on difficulties in the synthesis and merits of Maxwell dimensions.

Materials & methods

Supplementary to BEME systematic review 52, the co-authors of this paper analysed their independent and consensus classification of evidence from the 45 primary studies [17: 3,228 ‘initial hits’ on 1988–2016 bibliographic database search]. This involved much immersion and deliberation on alternative interpretations, within the pragmatism paradigm [25]. These articles presented K2, K3, or K4 +/-K1 evidence, as K1-only articles had been excluded. ‘Self-reported’ evidence was allowed. For that review, a combined deductive content analysis and thematic analysis [3,26,27] focused on integrative synthesis to summarize the evidence systematically, quantifying as appropriate [28]. QSR NVivo 10 assisted data handling.

After calibrating 10 articles together, each reviewer also coded evidence in each article to one or more:

- K1-K4 and
- Maxwell dimensions of quality adapted to the mobile device (left column, Table 1): *acceptability, accessibility, appropriateness, effectiveness, efficiency, and equity* (expanded to equity-ethics-professionalism).

Disagreements were resolved by discussion. Discussion of these classifications also informed the overall integrative synthesis.

For each Maxwell dimension, analysis in IBM SPSS Statistics 24 calculated the percentage of the articles where, respectively, the main body presented or the abstract suggested empirical evidence. Cohen kappa measured the reliability of the independent Yes/No classifications of the main body (Reviewer 1 vs Reviewer 2). Imbalance in average prevalence of Yes vs No (near the extremes rather than 50%) [29,30, p.260] prompted use of prevalence-adjusted, bias-adjusted kappa (PABA-K) with 95% confidence interval.

Hypothesis-testing compared Yes-No for each Maxwell dimension from:

- combined pre-agreement (i.e. Yes = both independently coded to Yes; No = one or both coded to No) vs final agreed classification (post-discussion consensus) of main body of article.
- abstract vs main body.

McNemar hypothesis-testing treated both these comparisons as paired, under the null hypothesis of no difference in Yes-No.

SPSS 24 hierarchical cluster analysis explored how Maxwell dimensions and K-levels overlapped. An initial basic, simple-linkage, nearest-neighbour cluster analysis suggested the order for variables to enter a between-groups linkage analysis, clustering by variable (measuring Squared Euclidean distance; binary variable: 1 = Yes, 2 = No, rescaled to 0–1). A dendrogram summarized, from the left, stronger relationships with shorter horizontal fork-prongs, with vertical lines ‘joining’ variables (x-axis distance) (Figure 1) [31]. Coherence of findings and a ‘scree plot’ (agglomeration coefficient vs stage) suggested how many clusters to declare.

Narrative summary captured reflection on the difficulties in synthesis [28]. Both reviewers independently outlined four main difficulties in synthesis and four main merits in using Maxwell dimensions in synthesis and analysis, then agreed a final summary in reflective discussion.

Results

Distribution and reliability of using Maxwell dimensions and overlap with K-levels

All but one study provided evidence for appropriateness of mobile device use to learning needs (including caregiving and patient safety aspects) (44/45, 97.8%), 86.7% each for acceptability and accessibility, with 73.3% and 44.4% for effectiveness and efficiency, respectively, but only just over one-quarter for equity-ethics-professionalism (Table 1). For the main body of articles, independent observations of Maxwell dimensions agreed best for appropriateness (good, PABA-K = 0.78). Classifying to acceptability, accessibility, effectiveness, and equity-ethics-professionalism (mostly about digital professionalism) showed moderate agreement (PABA-K = 0.47–0.51). Classifying to efficiency reached only fair agreement (PABA-K = 0.33), broadly interpreted as consideration of outputs to inputs, such as costs, saving or making the most of time or effort in learning or providing care, or allowing timely feedback.

The largest cluster in the five-cluster dendrogram of K-levels and Maxwell dimensions was acceptability–accessibility–K1–appropriateness–K3, with effectiveness

Table 1. Distribution and reliability: Maxwell dimensions (and final clustering with Kirkpatrick K-levels noted). A systematic review (Maudsley et al. 2019): *What works best for health professions students using mobile (hand-held) devices for educational support on clinical placements? Evidence from n = 45 studies.*

| K-level in 5-cluster dendrogram | In: main body of paper: | | Pre-discussion (initial): <i>Two independent observations</i> | | In: abstract: | In: main body of paper VS abstract | In main body of paper: Present pre- (initial)* vs post-discussion | |
|---------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------|---------------------------------|---------------------------------|------------------------------------|-------------------------------------------------------------------|------------------------|
| | Present post-discussion no. (%) | Strength of agreement (95% confidence interval) Cohen K | Present post-discussion no. (%) | McNemar p | | | | Y: Yes Classified same |
| K1 | Maxwell dimension adapted to evidence re use of device | | | | | | | |
| | Acceptability what students prefer and how satisfied they are | 39 (86.7) | 0.16 | 0.51 (0.22 to 0.80) moderate | 30 (66.7) | 0.004 | 31 0 8 6 (82.2) | 0.008 |
| | Accessibility how much they use it, for what purposes/advantages, with what barriers | 39 (86.7) | 0.34 | 0.51 (0.22 to 0.80) moderate | 34 (75.6) | 0.063 | 29 0 10 6 (77.8) | 0.002 |
| K3 | Appropriateness how it meets challenges/learning needs in clinical practice and students' work relationships (with patients/peers/staff) and capabilities for care and safety | | 44 (97.8) | -0.06 | 0.78 (0.53 to 1.00) good | 0.063 | 39 1 5 0 (86.7) | 0.219 |
| | Effectiveness whether the device works (to improve learning on clinical placement) | 33 (73.3) | -0.04 | 0.47 (0.18 to 0.76) moderate | 28 (62.2) | 0.063 | 31 2 2 10 (91.1) | 1.000 |
| K4b | Efficiency how it affects outputs:inputs | 20 (44.4) | 0.36 | 0.33 (0.06 to 0.61) fair | 14 (31.1) | 0.031 | 14 1 6 24 (84.4) | 0.125 |
| K2a | K4a | Equity-ethics-professionalism how it relates to fairness (a dimension widened to include ethical and professionalism aspects of using the device) | 12 (26.7) | 0.12 | 0.51 (0.22 to 0.80) moderate | 0.250 | 1 0 11 33 (75.6) | 0.001 |
| K2b | | | | | | | | |

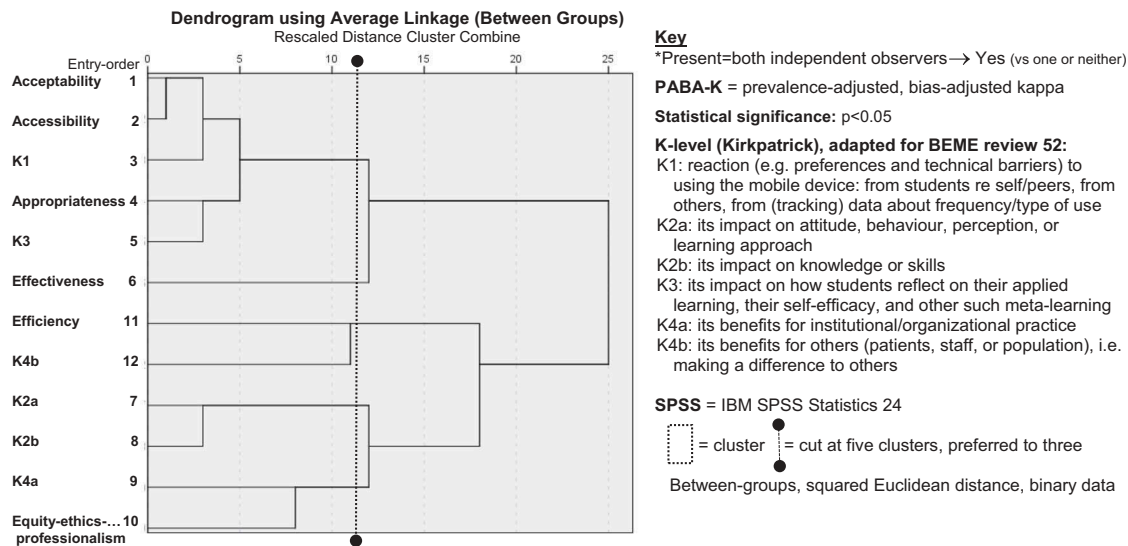


Figure 1. SPSS hierarchical five-cluster dendrogram: Maxwell dimensions and K-levels.

A systematic review [17]: *What works best for health professions students using mobile (hand-held) devices for educational support on clinical placements?* Evidence from $n = 45$ studies

nearby, quite dissimilar from the efficiency–K4b cluster (Figure 1). K2a–K2b clustered stronger than K4a–equity-ethics-professionalism.

How abstracts represented Maxwell quality dimensions of evidence

Acceptability was classified as ‘present’ significantly more post-discussion (same in 82.2%, $p = 0.008$; 31/45 → 39/45) and likewise for accessibility (same in 77.8%, $p = 0.002$; 29/45 → 39/45) and equity-ethics-professionalism (same in 75.6%; $p = 0.001$; 1/45 → 12/45).

If the abstract suggested evidence of a Maxwell dimension, the main body included that evidence. The main body presented evidence for acceptability and efficiency significantly more than the abstract: 86.7% vs 66.7%, $p = 0.004$; 44.4% vs 31.1%, $p = 0.031$, respectively. For accessibility, appropriateness, effectiveness ($p = 0.063$, respectively) and equity-ethics-professionalism ($p = 0.250$), the excess was not statistically significant.

Reflecting on difficulties in synthesis and analysis and merits of Maxwell dimensions

Combined observations about the main struggles in synthesis and analysis aggregated around the time (Figure 2):

- before: *insufficient guidance about such mixed evidence and concerns about transgressing qualitative research ‘rules’*,
- during: *ill-defined outcomes and methods, requiring much translation, particularly for equity-ethics-professionalism evidence*,

- after: *much effort in reporting analysis of the process and much potential to be misconstrued*.

Likewise, for the list of Maxwell dimensions, main merits aggregated around the time (Figure 2):

- before: *a simplified starting-point, adaptable to the intervention*,
- during: *a lens for better and wider understanding of implementation and impact, intended or otherwise*,
- after: *a structured framework for reflection, prompting deliberation to consensus about tricky evidence and its wider worth (and possible gaps), and being consistent with application of Maxwell dimensions by systematic reviews elsewhere in health services research*.

Discussion

Systematic reviews in medical education require more open discussion of difficulties in synthesizing suboptimal evidence-bases and more systematic reporting of process analysis. In the example explored here, Maxwell dimensions [22,23] helped to classify and synthesize the variegated quantitative, qualitative, and mixed methods evidence-base of a systematic review [17] meaningfully and moderately reliably, when only one-third of articles reported ‘quantitative-only’ research. The dimensions also helped to illuminate a coherent relationship with K-levels. Health professions education systematic reviews have not used this framework previously and very few health services research systematic reviews have reported using it [32–34]. As with the Kirkpatrick model, Maxwell dimensions were valuable in simplifying the approach

| |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>The main struggles encountered in synthesis were:</p> <p><u>Before starting</u></p> <ul style="list-style-type: none"> • There was little guidance on how to synthesize such disparate evidence from qualitative, quantitative, and mixed methods research. • There was a nagging concern that the qualitative research would either not be given due weight or the review approach would transgress some strict rule about synthesizing evidence out of context of the tradition and interpretation of the primary study. <p><u>When reviewing</u></p> <ul style="list-style-type: none"> ❖ Outcome measures were often poorly defined in the evidence-base. ❖ Authors' approaches to evaluation of the effects of an intervention were often unstructured or any structure appeared post hoc. ❖ Before Kirkpatrick or Maxwell coding could be negotiated and agreed, much 'reading between the lines' was required of the reviewers to interpret outcomes. ❖ It was relatively unusual for authors to give sufficient information to allow confident coding (or not) to the 'equity-ethics-professionalism' dimension. <p><u>When reporting the review</u></p> <ul style="list-style-type: none"> • Articulating and analysing what was done in synthesis was challenging, despite the reviewers having a shared understanding. • Despite working within the pragmatism paradigm, both reviewers were concerned that reporting the quantitative analysis of the studies and of the process would be misconstrued as suggesting a flawed or overzealous approach, despite its being bespoke and appropriate for that specific evidence-base. |
| <p>The main merits of using Maxwell dimensions in synthesis and analysis were:</p> <p><u>Before starting</u></p> <ul style="list-style-type: none"> • They were easily adapted to the specific educational intervention (the mobile device on clinical placement). • They gave a simplified starting-point when both reviewers were floundering about what to do next, given that the write-ups of other people's systematic reviews sounded much more straightforward, even if they were not. <p><u>When reviewing</u></p> <ul style="list-style-type: none"> ❖ They gave a clearer view of the impact of an intervention on the participants. ❖ They made it easier to understand participant adherence with an intervention (use of mobile device). ❖ They gave broader understanding of the study outcomes, intended or otherwise. <p><u>When reporting the review</u></p> <ul style="list-style-type: none"> ❖ They gave a structured framework for reflection generally. • They prompted useful deliberation specifically about the nature and worth of various aspects of the evidence beyond 'Does it work?' and highlighted potential gaps. • Noting how other authors in health services research had used them as a framework for concepts in systematic reviews (in the search or in the analysis) increased both reviewers' confidence about the approach taken. |

Figure 2. Two reviewers' reflections (•❖) on struggles in synthesis and merits of Maxwell dimensions in synthesis and analysis.

A systematic review [17]: *What works best for health professions students using mobile (hand-held) devices for educational support on clinical placements?* Evidence from n = 45 studies

[35] plus giving a deliberative framework for reviewers to share understanding in a tricky integrative synthesis and consider possible gaps. While the Maxwell list of 'characteristics' might not amount to a 'framework' for health-care performance measurement and improvement [36], here Maxwell dimensions provided a 'framework for concepts' of educational support. This built beyond 'Does it work?'

When BEME review 52 [17] reported much potential for mobile devices to support health professions students on clinical placement (via their transitions, meta-learning, and care contribution, but requiring policy to address negative informal and hidden curricula), this explored beyond 'Does it work?'. Besides the usual recommendation for improved reporting of primary research, BEME

review 52 recommended that 'effectiveness'-reviews extend beyond a simplistic approach of just 'What works?', echoing Eva [37] and Regehr [24]. This justified further scrutiny. This also reflected the broader horizons of health services evaluation [38] and challenges in systematic reviews of complex interventions in public health and social sciences [39,40], including the need to synthesize haphazard evidence. The 'stainless steel' law of such systematic reviews remained that '*the more rigorous the review, the less evidence there will be to suggest that the intervention is effective*' [39, p.758]. In qualitative synthesis:

"study findings are systematically interpreted through a series of expert judgements to represent the meaning of the collected work. ... the findings of

qualitative studies – and sometimes mixed-methods and quantitative research – are pooled.” [3, p.253]

Better narrative synthesis is required [39] plus better blending with quantitative observations, as appropriate. Here, further analysis confirmed Maxwell dimensions to be a useful extra framework to prompt much deliberation on evidence, improve understanding, and represent complexity.

Maxwell dimensions typified health-care evaluation in the UK National Health Service (NHS) during its 1990s quality management and ‘internal’ market phase [41], providing a ‘characteristics model’ of quality [42]. In health services research, these dimensions have guided the integrative systematic review of ‘*What is the effect of non-medical prescribing in primary care and community settings on patient outcomes?*’, exploring beyond ‘*Does it work?*’ [32]. A similar US Institute of Medicine [43] list of ‘characteristics’ (explicitly mentioning patient safety) was also popular, guiding classification of evidence in a systematic review of pay-for-performance in UK general practice [33]. While Donabedian’s much revered structure-process-outcome framework for measuring health-care quality might have been an alternative [44], it did not intuitively have as much potential for exploring the K-levels. Berwick and Fox [45] considered that the Donabedian framework was not necessarily patient-centred or viewing health care as a system. The Maxwell dimensions allowed the evidence in the educational context to be student-centred and to be viewed holistically as if part of a learning system. Several major NHS reforms and many alternative quality indicators later, more complex representations of quality have superseded Maxwell dimensions, e.g. to analyse patient-professional co-production of knowledge and health [46].

Nevertheless, Maxwell dimensions remain a basic, durable, practical starting-point for evaluating services [e.g. 47,48], notably dental in recent years [e.g. 49,50]. Relevant to workforce development, Halter et al.’s [34] systematic review of the impact of physician associates on secondary care used Maxwell dimensions as ‘outcome’ search-terms and then reportedly to organize the main messages (albeit without presenting or discussing the latter). Here, the dimensions adapted well to making sense of a research evidence mash-up about mobile devices in clinical placements for the future health professions workforce.

While the evidence is not obliged to represent all dimensions, the extra Maxwell lens highlighted potentially underrepresented aspects such as efficiency and equity-ethics-professionalism and an evidence distribution more towards the 3As. There was a gap in the evidence about quantifying efficiencies in learning or care provided. There was also a gap in the evidence about ‘fairness’ of the use of mobile devices.

Abstracts tended to omit much supplementary evidence about acceptability understandably (given

exclusion of K1-only papers) but omitting evidence about efficiency suggested lower sensitivity of title-abstract filtering on this dimension. Despite moderately reliable independent observations, inter-observer discussion significantly increased the proportion classified to acceptability, accessibility, and especially equity-ethics-professionalism, probably reflecting refinement of definitions but also the subtlety of some evidence. A coherent clustering with K-levels confirmed that broadly interpreting ‘effectiveness’ across research types reached well beyond K2a-K2b randomized controlled trial-type evidence. Widening impact to ‘making a difference’ for the organization or for other people involved K4a and K4b clustering with, respectively, equity-ethics-professionalism and efficiency – and not with K1, so even if K-levels were non-hierarchical, this suggested that K1 and K4a/b differ substantively.

Yardley and Dornan [8] found ‘K2 and below’ to show suboptimal sensitivity as a BEME exclusion-filter for their review-question (about early workplace experience in undergraduate medical education). For BEME review 52 though, excluded K1-only papers did not illuminate its review-question further. Suitability of K-levels to filter and K-levels and Maxwell dimensions to organize and summarize trustworthy evidence depends on the review-question.

Strengths here were that both Kirkpatrick model and Maxwell dimensions were applied with critical scepticism and within the pragmatism paradigm, which enhanced: deliberation and synthesis; quantification of key aspects (Table 1, Figure 1); and mixing of qualitative and quantitative analytical approaches [51]. Furthermore, despite problematic calibration of Maxwell coding for equity-ethics-professionalism, the final coding appeared robust.

The evidence-base on which this ‘process analysis’ focused was relatively small, yet its eclecticism was both strength and weakness. Calibration about efficiency needed more attention. It was also unsurprising if evidence about acceptability and accessibility was uncommon, given exclusion of K1-only articles, but useful K1 supplementary evidence still featured. While using cluster analysis on 12 variables for $n = 45$ ignored a 2^m ‘rule-of-thumb’ sample-size (where $m =$ number of variables) [31] and might be seen as overkill (or overreliant on hypothesis-testing [40]), the five-cluster dendrogram illuminated the Kirkpatrick–Maxwell relationship. Exploration of Maxwell coding against strength of evidence and study design may well also be merited but would require a larger evidence-base.

Conclusion

Beyond the convenience of Kirkpatrick outcome-levels for filtering abstracts and summarizing outcome-evidence, Maxwell dimensions helped to promote Regehr’s [24] preferred imperatives for medical

education evidence: *gaining a rich understanding and representing complexity*. This contrasted with what he called the dominant imperatives: *seeking proof* [*that something works*, 37, p. 295] and *generalizable simplicity*.

Reviewer-pairs or teams must calibrate and critique such classification tools with care to widen analytical horizons robustly. Here, deliberative synthesis of ‘whether’, ‘how/why’, and ‘what’ concepts [2,18,52,53] applied Maxwell dimensions to illuminate the process of implementing the educational intervention (using mobile devices on clinical placements) as well as broadly interpreting outcomes. To improve systematicity [54,55] and thoroughness for tricky integrative synthesis in systematic reviews, reflective deliberation and supplementary analyses about such tools are required, particularly their conceptual integrity and trustworthiness. Maxwell dimensions at least give a practical framework for organizing, deliberating about, and synthesizing key concepts when struggling beyond Kirkpatrick in a messy evidence-base. Waiting for the perfect evidence-base to synthesize would be unhelpful for the topic. As Glass [56, p.4] highlighted:

“A common method of integrating several studies with inconsistent findings is to carp on the design or analysis deficiencies of all but a few studies – those remaining frequently being one’s own work or that of one’s students or friends – and then advance the one or two ‘acceptable’ studies as the truth of the matter. This approach takes design and analysis too seriously, in my opinion. I don’t condone a poor job of either; but I also recognize that a study with a half dozen design and analysis flaws may still be valid. [...] ... I believe the difference [in results between poorly-designed and the best-designed studies] to be so small that to integrate research results by eliminating the ‘poorly done’ studies is to discard a vast amount of important data.”

Wilson and Lipsey [57, p.420] confirmed this in analysing 250 meta-analyses:

“It appears that low method quality functions more as error than as bias, reducing the confidence that can be placed in the findings but neither consistently over- nor underestimating program effects.”

In health professions education, it is challenging to undertake systematic multi-component mixed methods reviews that attempt to arrange, interpret, and summarize evidence about *‘What is the effect of this complex intervention?’* [58, p.2]. Such reviews may well attempt to ‘configure’ and (less so) ‘aggregate’ [58] an evidence-base that is quite a mess, epistemologically or otherwise. More tools and analysis of their use are required for the synthesis of jumbles of evidence.

Practice points

- There should be more systematic reporting of process analysis about difficulties synthesizing suboptimal evidence-bases in health professions education.
- Maxwell dimensions are potentially useful for evaluating educational interventions and synthesizing messy, variegated evidence-bases.
- Kirkpatrick model and Maxwell dimensions applied to such an evidence-base clustered coherently together and should be applied with critical scepticism to improve understanding and represent complexity.
- To improve systematicity and thoroughness of systematic reviews, especially when synthesis is tricky, reflective deliberation and supplementary analyses about the conceptual integrity and trustworthiness of ‘filtering and classification’ tools are warranted.
- Be aware that abstracts may well omit certain types of substantive evidence reported in the main body, thus reducing the sensitivity of title-abstract filtering.

Acknowledgments

We would like to thank our other BEME review 52 co-authors for their contribution to that systematic review on which this process analysis built.

Small extracts of an early version of this work were presented at the Association for Medical Education in Europe (AMEE) conference in Helsinki, Aug-2017 (poster).

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

David Taylor  <http://orcid.org/0000-0002-3296-2963>

References

- [1] Thistlethwaite J, Davies H, Dornan T, et al. What is evidence? Reflections on the AMEE symposium, Vienna, August 2011. *Med Teach.* 2012;34(6):454–457. Epub 2012 Apr 11.
- [2] Gordon M. Are we talking the same paradigm? Considering methodological choices in health education systematic review. *Med Teach.* 2016 Jul;38(7):746–750.
- [3] Bearman M, Dawson P. Qualitative synthesis and systematic review in health professions education. *Med Educ.* 2013;47(3):252–260.
- [4] Denzin NK. The elephant in the living room: or extending the conversation about the politics of evidence. *Qual Res.* 2009;9(2):139–160.

- [5] Dixon-Woods M, Bonas S, Booth A, et al. How can systematic reviews incorporate qualitative research? A critical perspective. *Qual Res.* 2006;6(1):27–44.
- [6] Kirkpatrick D. Great ideas revisited: techniques for evaluating training programs. Revisiting Kirkpatrick's four-level model. *Training Dev.* 1996 Jan;50(1):54–59.
- [7] Frye AW, Hemmer PA. Program evaluation models and related theories: AMEE Guide No. 67. *Med Teach.* 2012;34(5):e288–e299.
- [8] Yardley S, Dornan T. Kirkpatrick's levels and education evidence. *Med Educ.* 2012;46(1):97–106.
- [9] Campbell K, Taylor V, Douglas S. Effectiveness of online cancer education for nurses and allied health professionals: a systematic review using Kirkpatrick Evaluation Framework. *J Cancer Educ.* 2019;34(2):339–356.
- [10] Gorbanev I, Agudelo-Londoño S, González RA, et al. A systematic review of serious games in medical education: quality of evidence and pedagogical strategy. *Med Educ Online.* 2018;23(1):1–9.
- [11] Johnston S, Coyer FM, Nash R. Kirkpatrick's evaluation of simulation and debriefing in health care education: a systematic review. *J Nurs Educ.* 2018;57(7):393–398.
- [12] Moreau KA. Has the new Kirkpatrick generation built a better hammer for our evaluation toolbox? *Med Teach.* 2017;39(9):999–1001.
- [13] Kirkpatrick JD, Kirkpatrick WK. Kirkpatrick's four levels of training evaluation. Alexandria (VA): ATD Press; 2016.
- [14] Lechner SK. Evaluation of teaching and learning strategies. *Med Educ Online.* 2001;6:1–4.
- [15] Gordon M, Carneiro AV, Patricio MF. Enhancing the impact of BEME systematic reviews on educational practice [letter]. *Med Teach.* 2015;37(8):789–790.
- [16] Petticrew M. Time to rethink the systematic review catechism? Moving from 'what works' to 'what happens'. *Syst Rev.* 2015;4(36):1–6.
- [17] Maudsley G, Taylor DCM, Allam O, et al. A Best Evidence Medical Education (BEME) systematic review of: what works best for health professions students using mobile (hand-held) devices for educational support on clinical placements? BEME Guide No. 52. *Med Teach.* 2019 Feb;41(2):125–140.
- [18] Cook DA, Bordage G, Schmidt HG. Description, justification and clarification: a framework for classifying the purposes of research in medical education. *Med Educ.* 2008 Feb;42(2):128–133.
- [19] Colthart I, Bagnall G, Evans A, et al. 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(2):124–145.
- [20] Hammick M, Dornan T, Steinert Y. Conducting a best evidence systematic review. Part 1: from idea to data coding: BEME Guide No. 13. *Med Teach.* 2010;32(1):3–15.
- [21] Harden RM, Grant J, Buckley G, et al. BEME Guide No. 1: best evidence medical education. *Med Teach.* 1999;21(6):553–562.
- [22] Maxwell RJ. Quality assessment in health. *Br Med J.* 1984;288:1470–1472.
- [23] Maxwell R. Dimensions of quality revisited: from thought to action. *Qual Health Care.* 1992;1(3):171–177.
- [24] Regehr G. It's NOT rocket science: rethinking our metaphors for research in health professions education. *Med Educ.* 2010;44(1):31–39.
- [25] Creswell JW. Research design: qualitative, quantitative, and mixed methods approaches. 2nd ed. London: Sage Publications; 2003.
- [26] Green J, Thorogood N. Qualitative methods for health research. 3rd ed. London: Sage Publications; 2014.
- [27] Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. *Nurs Health Sci.* 2013 Sep;15(3):398–405. Epub 2013 Mar 11.
- [28] Dixon-Woods M, Agarwal S, Jones D, et al. Synthesising qualitative and quantitative evidence: a review of possible methods. *J Health Serv Res Policy.* 2005;10(1):45–53.
- [29] Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol.* 1993 May;46(5):423–429.
- [30] Looney SW, Hagan JL. Chapter 5: validation of biomarkers. In: Analysis of biomarker data: A practical guide. Hoboken (NJ): John Wiley & Sons; 2015. p. 255–331.
- [31] Sarstedt M, Mooi E. Chapter 9: cluster analysis. In: A concise guide to market research. Springer Texts in Business and Economics. 3rd ed. Berlin: Springer-Verlag; 2019. p. 301–354.
- [32] Bhanbhro S, Drennan VM, Grant R, et al. Assessing the contribution of prescribing in primary care by nurses and professionals allied to medicine: a systematic review of literature. *BMC Health Serv Res.* 2011;11:10.
- [33] Gillam SJ, Siriwardena AN, Steel N. Pay-for-performance in the United Kingdom: impact of the Quality and Outcomes Framework: a systematic review. *Ann Fam Med.* 2012 Sep–Oct;10(5):461–468.
- [34] Halter M, Wheeler C, Pelone F, et al. Contribution of physician assistants/associates to secondary care: a systematic review. *BMJ Open.* 2018 Jun 19;8(6):e019573:1–21.
- [35] Reio TG, Rocco TS, Smith DH, et al. A critique of Kirkpatrick's evaluation model. *New Horiz Adult Educ Hum Resour Dev.* 2017;29(2):35–53.
- [36] Klassen A, Miller A, Anderson N, et al. Performance measurement and improvement frameworks in health, education and social services systems: a systematic review. *Int J Qual Health Care.* 2010;22(1):44–69.
- [37] Eva KW. Broadening the debate about quality in medical education research. *Med Educ.* 2009 Apr;43(4):294–296.
- [38] Maudsley G. What issues are raised by evaluating problem-based undergraduate medical curricula? Making healthy connections across the literature. *J Eval Clin Pract.* 2001;7(3):311–324.
- [39] Petticrew M. Why certain systematic reviews reach uncertain conclusions. *Br Med J.* 2003;326(7392):756–758.
- [40] Threlfall AG, Meah S, Fischer AJ, et al. The appraisal of public health interventions: the use of theory. *J Public Health.* 2015;37(1):166–171.
- [41] Taylor D. Quality and professionalism in health care: a review of current initiatives in the NHS. *Br Med J.* 1996;312(7031):626–629.
- [42] Raven JH, Tolhurst RJ, Tang SL, et al. What is quality in maternal and neonatal health care? *Midwifery.* 2012;28(5):E676–E683.
- [43] Institute of Medicine (US National Academy of Sciences) Committee on Quality of Health Care in America; Richardson WC, Berwick DM, Bisgard JC, et al. Crossing the quality chasm: a new health system for the 21st Century [brief report]. Washington (DC): National Academies Press; 2001. p. 1–8. [cited 2020 Feb]; Available from: <http://www.nationalacademies.org/hmd/~/media/Files/Report%20Files/2001/Crossing-the-Quality-Chasm/Quality%20Chasm%202001%20%20report%20brief.pdf>
- [44] Donabedian A. Evaluating the quality of medical care. *Milbank Mem Fund Q.* 1966;44(3):166–206. Reprinted: *Milbank Q.* 2005;83(4):691–729.

- [45] Berwick D, Fox DM. “Evaluating the quality of medical care”: Donabedian’s classic article 50 years later. *Milbank Q.* 2016;94(2):237–241.
- [46] Mulley A, Coulter A, Wolpert M, et al. New approaches to measurement and management for high integrity health systems. *Br Med J.* 2017 Mar;356:j1401:1–7.
- [47] Harries U, Landes R, Popay J. Visual disability among older-people: a case-study in assessing needs and examining services. *J Public Health Med.* 1994;16(2):211–218.
- [48] Rajpura A, Sethi S, Taylor M. An evaluation of two rapid access chest pain clinics in central Lancashire, UK. *J Eval Clin Pract.* 2007 Jun;13(3):326–336.
- [49] Chideka K, Klass C, Dunne S, et al. Listening to older adults: community consultation on a new dental service. *Community Dent Health J.* 2015 Dec;32(4):231–236.
- [50] Worsley DJ, Marshman Z, Robinson PG, et al. Evaluation of the telephone and clinical NHS urgent dental service in Sheffield. *Community Dent Health.* 2016 Mar;33(1):9–14.
- [51] Maudsley G. Mixing it but not mixed-up: mixed methods research in medical education (a critical narrative review) [Web-paper]. *Med Teach.* 2011;33(2):e92–e104.
- [52] Gordon M, Darbyshire D, Baker P. Separating the wheat from the chaff: the role of systematic review in medical education. *Med Educ.* 2013 Jun;47(6):632.
- [53] Gordon M, Vaz Carneiro A, Patricio M, et al. Missed opportunities in health care education evidence synthesis [letter]. *Med Educ.* 2014 Jun;48(6):644–645.
- [54] Eva KW. On the limits of systematicity. *Med Educ.* 2008;42(9):852–853.
- [55] Gordon M, Daniel M, Patricio M. What do we mean by ‘systematic’ in health education systematic reviews and why it matters! *Med Teach.* 2019 Aug;41(8):956–957.
- [56] Glass GV. Primary, secondary, and meta-analysis of research. *Educ Researcher.* 1976;5(10):3–8.
- [57] Wilson DB, Lipsey MW. The role of method in treatment effectiveness research: evidence from meta-analysis. *Psychol Methods.* 2001;6(4):413–429.
- [58] Gough D, Thomas J, Oliver S. Clarifying differences between review designs and methods. *Syst Rev.* 2012 Jun 9;1(28):1–9.