



Reviews of teaching methods – which fundamental issues are identified?

Åsa Hirsh, Claes Nilholm, Henrik Roman, Eva Forsberg & Daniel Sundberg

To cite this article: Åsa Hirsh, Claes Nilholm, Henrik Roman, Eva Forsberg & Daniel Sundberg (2022) Reviews of teaching methods – which fundamental issues are identified?, Education Inquiry, 13:1, 1-20, DOI: [10.1080/20004508.2020.1839232](https://doi.org/10.1080/20004508.2020.1839232)

To link to this article: <https://doi.org/10.1080/20004508.2020.1839232>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 15 Nov 2020.



[Submit your article to this journal](#)



Article views: 11589



[View related articles](#)



[View Crossmark data](#)



Citing articles: 2 [View citing articles](#)

ARTICLE



Reviews of teaching methods – which fundamental issues are identified?

Åsa Hirsh^a, Claes Nilholm^{ib}, Henrik Roman^b, Eva Forsberg^{ib}
and Daniel Sundberg^{ib}^c

^aDepartment of Education and Communication, Jönköping University, Jönköping, Sweden; ^bUppsala University, Uppsala, Sweden; ^cLinnaeus University, Kalmar, Sweden

ABSTRACT

The purpose of this study is to discern and discuss issues with relevance to the tension between contextuality and generalisation, which recurrently are identified over time in research reviews of teaching methods. The 75 most cited reviews on teaching methods listed in the Web of Science from 1980 to 2017 were analysed. Since our interest is the claims made in each article about the teaching method under study, the analysis concerned the abstract, results, discussion, conclusion, and implication parts of each review. Three main issues, cutting across the reviews over time, were identified: 1) the abundance of moderating factors, 2) the need for highly qualified teachers, and 3) the research-practice gap. It is argued that the three issues reflect tensions in original research. The implications of these findings are discussed in the article. One main conclusion is that such issues ought to be more explicitly attended to and elaborated in both primary and secondary level research. The importance of viewing validity as a multidimensional concept, including internal, external, and ecological aspects, is underlined. Further, ideas from realistic reviewing are used to discuss a contextually bound approach to causality.

ARTICLE HISTORY

KEYWORDS

Internal and external validity; moderating factors; overview; research-practice gap; review; teaching methods

Introduction

Given the global emphasis on education as a road to national and individual success, it is not surprising that a vast amount of research concerns which teaching methods enable education to fulfil its aims. Although education concerns many areas, such as educational policy, the organisation of education, financial systems, and school leadership, there seems to be wide agreement that teaching, in the end, is the key factor in making educational systems successful (e.g. Barber & Mourshed, 2007; Hargreaves & Fullan, 2012; Hattie, 2003; OECD, 2016; Stigler & Hiebert, 2009).

How teaching should be arranged in the best possible way has been targeted in a great number of investigations involving different theoretical points of departure (Hattie, 2009). Consequently, reviews of the effectiveness or appropriateness of teaching methods have become increasingly available. Producing such reviews is a logical way to integrate findings and insights from different studies. Systematic research reviews can

CONTACT Åsa Hirsh ✉ asa.hirsh@ju.se 📍 Jönköping University, Jönköping, Sweden

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

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

contribute in various ways with knowledge that may inform research, practice and policy decisions (cf. Gough, Thomas, & Oliver, 2012). However, findings from underlying studies often show mixed and sometimes even conflicting results, due to a variety of factors (e.g. Shute, 2008). Instructional methods and interventions act in complex systems, and their effects are dependent on various factors in the context as well as the ways in which and by whom they are implemented and enacted (cf. Cartwright & Hardie, 2012; Pawson, 2006; Pawson, Greenhalgh, Harvey, & Walshe, 2005; Rycroft-Malone et al., 2012).

Synthesising the results and effects of numerous primary studies inevitably involves a certain degree of decontextualization. Still, at secondary research level, researchers recognise, relate to, and/or problematise the meaning and impact of the context in various ways. In the current study, we develop knowledge on how the tension between contextuality and generalisability is addressed and elaborated in research reviews of teaching methods. We are particularly interested in whether and how issues concerning what works for whom and in what circumstances are problematised (cf. Pawson, 2006). Thus, we explore those issues that recur across studied methods and overtime in research reviews of teaching methods, with relevance to the tension between context and generalisation. Subsequently, identified issues will be discussed in terms of possible implications for both primary and secondary level research.

The overview format

This study is situated within the frames of a research project with the overall aim of increasing and refining our knowledge about teaching and teaching research (Hirsh & Nilholm, 2019; Roman, Sundberg, Hirsh, Nilholm, & Forsberg, 2018). In order to clarify the context in which the present study has emerged, a brief description of starting points and assumptions driving the overall research project follows below (cf. Nilholm & Göransson, 2017).

With a growing amount of primary research, the number of research reviews, i.e. a secondary level that sums up and synthesises primary level research on a particular topic, has also increased. This in turn paves the way for a tertiary level of research that summarises and synthesises research reviews (Polanin, Maynard, & Dell, 2017). Various terms are used to describe the type of third-order research that uses research reviews as its empirical data, such as overview (Polanin et al., 2017), meta-meta-analysis (Hattie, 2009; Kazrin, Durac, & Agteros, 1979), meta-synthesis (Cobb, Lehmann, Newman-Gonchar, & Alwell, 2009), review of reviews (Maag, 2006), tertiary review (Torgerson, 2007), mega-analysis (Terhart, 2011) and umbrella review (Grant & Booth, 2009). The terms used thus vary, but common is that the tertiary level is targeted at synthesising the evidence on a particular topic of interest by examining only the highest level of evidence, i.e. the evidence presented in systematic reviews or meta-analyses. In our project, we have chosen to use the term overview.

Characteristic of our overview methodology is, among other things, the selection of research reviews to be included. The point of departure is that it is fruitful to map and analyse the research that the research community itself considers important (cf. Nilholm & Göransson, 2017). Therefore, we only include recognised, high-impact and top-cited research listed in the Web of Science (WoS).¹ An overall interest in the

project is to explore trends in influential research reviews on teaching methods and to discern common findings and topics to discuss *across* issues – using the WoS as the influence indicator.

Other main differences between our overview methodology and methodologies focused on extracting evidence (see above) are that our type of overview a) is inclusive regarding different kinds of second-level review research methodologies, b) is inclusive with regard to different theoretical approaches at the review level (including reviews from critical interpretivist perspectives), and c) maps and analyzes several aspects of a research field, such as topics studied, theoretical/conceptual points of departure, methodologies used, and results and conclusions presented. Yet, one particular paper cannot elaborate on all of these aspects, and the present study concentrates on the results and conclusions presented, with special attention paid to the tension between contextuality and generalisability.

Method

Before the analysis specific to the present study could be carried out, extensive basic work had already been done, where the research group as a first step identified the 75 most cited research reviews on teaching methods listed in the WoS between 1980 and 2017 (25 from 1980 to 1999, 25 from 2000 to 2009, and 25 from 2010 to 2017).

Identifying reviews: search strategy and criteria for inclusion/exclusion

An initial search in the WoS Core Collection was carried out, using the simple search string: “teach* OR instr* OR curric* OR did* OR coach* OR guid* OR tut*,” restricted to topics within reviews in the four content areas (*education educational research, psychology education, education scientific disciplines, and education special*) during the time spans 1980–1989 (192 hits), 1990–1999 (446 hits), 2000–2009 (1241 hits), and 2010–2017 (2439 hits).

A second search in the WoS Core Collection was then carried out using the combined search string “teach* OR instr* OR curric* OR did* OR coach* OR guid* OR tut*” AND “review* or meta-analys* OR meta-narrative* OR meta-synthes* OR overview*,” restricted to topics within articles and reviews in the four content areas listed above during the time spans 1980–1989 (166 hits), 1990–1999 (1915 hits), 2000–2009 (3788 hits), and 2010–2017 (13,795 hits).

The result lists from the two searches were matched, and through the reading of abstracts, relevant top-cited reviews from each decade were identified and selected for further reading and coding. Our main inclusion criterion was that the review should focus on teaching methods in the K-12 context. Reviews concerned *only* with higher and/or adult education² and reviews on teacher conditional factors (educational background, class, gender, ethnicity, etc.) were excluded. Furthermore, we excluded reviews on learning which were not specifically related to classrooms, teaching, or school subjects, as well as reviews on societal aspects of school and schooling in a wider sense (i.e. which were not concerned with didactical matters). In addition, we excluded reviews on student and teacher wellbeing and self-perception/self-efficacy, unless they related to specific school interventions or teaching programmes. We did, however,

include a number of reviews on the use of technological artefacts for instructional purposes that to some degree also included external learning environments (outside classrooms and schools). For a full list of included reviews, see Appendix A.

Study feature coding

The 75 reviews were carefully read in their entirety and encoded in a 25-feature scheme (see [Appendix B](#)) developed by the research group. Twenty-five reviews were read and coded by two researchers in the group to check for interrater reliability, resulting in a 92% compliance. The coding of the remaining 50 reviews was divided between the two researchers. Different features of the coding scheme have been or are currently being used for different analyses in the various studies that are all part of the overall research project.

The current study

During the careful reading that the coding process entailed, it became apparent that underlying reviews largely discussed similar issues, drew similar conclusions, and/or pointed to similar implications for practice and/or research, regardless of teaching method studied. This, in turn, led to further analysis, guided by an overall interest in inductively and more deeply exploring the issues that appeared most frequently, with the aim to identify recurring issues and bring patterns of issues together in categories (cf. Saini & Shlonsky, 2012; Thomas & Harden, 2008).

Material in the current study

As a first step in the analysis, relevant features from the coding schemes of all the underlying 75 reviews were summarised in a table with the following captions:

The actual analysis in the current study concerned the summaries in the third through fifth columns of the table.

In the analysis phase, the summaries were regarded as text extracts that were the subject of qualitative content analysis. Content analysis is a flexible method for analysing text data obtained in various ways, such as interviews, observations, open-ended survey questions, or print media such as various types of articles, books, or policy documents (Cavanagh, 1997; Kondracki & Wellman, 2002). The goal of content analysis is “to provide knowledge and understanding of the phenomenon under study” (Downe-Wamboldt, 1992, p. 314), through systematic coding and identification of patterns (Hsieh & Shannon, 2005). Whatever type of text the content analysis takes its starting point in, the analysis starts at the manifest level. It may then proceed to the latent level, but not necessarily. The manifest analysis deals with the content aspect and describes the visible, obvious components (Downe-Wamboldt, 1992; Kondracki et al.,

2002), whereas the latent analysis deals with underlying meanings of the text (Downe-Wamboldt, 1992; Kondracki & Wellman, 2002).

When summarising results and implications of each of the reviews in the original coding process, our pronounced endeavour was to do so on a manifest level, that is, with as little abstraction or interpretation as possible. In all cases where possible, we took our starting point in the abstracts of the reviews, according to the logic that summarised there is what the authors themselves consider to be the most important results and implications. However, the results, discussion, conclusion, and/or implication parts of each review were also read in full, resulting in complementary text and more informative summaries than the very short lines appearing in the article abstracts.

Analysis in the current study

The following inductive analysis of the summaries was guided by Graneheim and Lundman's (2004) qualitative content analysis approach and conceptualisation. Each summary of results, implications for practice, and implications for research was considered a meaning unit that was labelled with codes. Coffey and Atkinson (1996, p. 32) describe codes as “tools to think with” and “heuristic devices” allowing the data to be thought of in different and/or new ways (while principally remaining on manifest data level). Since a code is assigned to discrete objects/phenomena, and each meaning unit often summarised various phenomena, each meaning unit was (in almost all cases) assigned more than one code.³

In a next step, the most frequently occurring codes were sorted into areas sharing a commonality, subsequently resulting in three overarching types of issues that we labelled as follows:

- (1) The abundance of moderating factors
- (2) The need for highly qualified teachers
- (3) The research–practice gap

Of these, categories 1 and 3 were so complex that further sorting into subcategories was carried out.

In the following results chapter, we use the term *overview findings* for our overarching categories (with associated subcategories). An overview finding can be described as a product of an accumulated analysis of individual review findings describing a phenomenon or aspects of a phenomenon (here teaching methods) (cf. Lewin et al., 2015). Overview findings thus arise in the analysis and involve interpretation. In general, overview findings can be formulated at different abstraction levels, depending on the degree of interpretation being made. Overview findings at a lower level of abstraction are often relatively close to underlying studies and formulated with concepts retrieved directly from them, while findings at a higher abstraction level may require other terms to be used.

Further, our analysis was partly⁴ guided by the methodology in the framework CERQual (which stands for *confidence in the evidence from reviews of qualitative research*) described by Lewin et al. (2015). A core purpose of CERQual is to offer a method for systematically and transparently assessing the weight (in terms of

coherence) of findings derived from qualitative research.⁵ Although our primary interest lies in describing recurrent patterns and in conducting a problematising discussion about those patterns, we acknowledge the importance of visualising the occurrence and frequency of different aspects (that together form our overview findings) in the various underlying reviews as a signal of the “weight” (in terms of coherence) of the overview findings. For this reason, we have created two tables highlighting the occurrence of specific aspects in the various included studies (see [Appendices C1](#) and [C2](#)).

Results

Before presenting the main results of the current study, i.e. the three overview findings, some overall observations are briefly accounted for regarding the format of the underlying reviews, as well as their temporal and geographical distribution.

The included 75 reviews build on different types of data in the primary studies, which largely affect the format of the reviews. Quantitative reviews, which are based on quantitative underlying studies, make up almost half of the sample (35/75). 24/75 reviews in the sample report both quantitative and qualitative data, whereas 16 reviews are explicitly qualitative. The distribution between the three different types of reviews is relatively even over the three periods 1980–1999, 2000–2009, and 2010–2017 (Roman et al., 2018).

The tables in [Appendices C1](#) and [C2](#) visualise the occurrence and frequency of different aspects in the underlying material. There, the reader can see which reviews elaborate on which aspects, which year the reviews were published, and the geographical distribution of the reviews in terms of national affiliations of the review authors. The Web of Science is located in the US, and there is a clear North American domination when it comes to the national affiliations of the authors. Three-quarters of all authors are affiliated with institutions in the US or Canada. The final quarter are affiliated with institutions in nine other countries: the Netherlands, the UK, Germany, Greece, Taiwan, Israel, Hong Kong, Australia, and Brazil.

The overview findings

As mentioned, in the analysis we found three overarching issues which were particularly frequently elaborated and discussed across the reviews. Below, we present these as overview findings, supported by typical excerpts.⁶

Overview finding 1: the abundance of moderating factors

Common to most of the research reviews is that they study the correlation between two variables, in the language of meta-analysis sometimes referred to as treatment and treatment outcome. The treatment variable is the teaching method, whereas the treatment outcome is the effect/impact of the teaching method on students’ learning/development. The relationship between treatment and treatment outcome is rarely or never direct, however. It is, obviously, dependent on the influence of so-called moderators, that is, variables that affect the strength of the relation between treatment and treatment outcome.

Although a more or less explicitly stated goal in several of the reviews is to give some kind of general answer concerning the impact of a given method, the reservations are ultimately many. Givers (teachers) as well as receivers (students) of the treatment are heterogeneous groups in several ways, and, additionally, there is great variation concerning the contextual conditions framing the teaching-learning process. Many moderators or combinations of moderators may potentially affect the method's impact on students' learning outcome.

This fact is also problematised and discussed in several of the included reviews that together constitute the empirical material underlying this study. An overview finding where coherence, thus, is strong (i.e. where a pattern is found across most of the underlying studies) is that a particular method has little or no effect *per se*; rather, our analysis shows that the effect depends on moderators linked to four (often inter-related) aspects (Table 1).

In an excerpt typical for many underlying reviews, Graham and Hebert (2011) conclude the following:

Just because a writing intervention was effective in improving students' reading in the studies included in this review does not guarantee that it will be effective in all other situations. In fact, there is rarely an exact match between the conditions in which the research was implemented and the conditions in which it is subsequently implemented by teachers. Mismatches between the conditions where a practice is implemented by a teacher and its effectiveness as established by researchers can vary widely, including differences

Table 1. Differences moderating outcomes of teaching methods.

Differences in students	<ul style="list-style-type: none"> - Age/grade level - Gender - Ethnicity/cultural identity - Achievement level - Cognitive level - Metacognitive level - Presence/degree of special needs - Beliefs/attitudes, motivation - Level of previous familiarity with method/artefact in focus
Differences in teachers	<ul style="list-style-type: none"> - Competence to see students' differences and needs - Competence to determine when and to what extent it is appropriate to use a method - Competence to design instruction in terms of tasks and classroom environment - Competence to see when explicit teaching is necessary/what level of involvement is appropriate in the teaching process - Competence to teach learning strategy skills (to work with students on the metalevel) - Relationship-building competence - Level of professional experience - Level of subject specific knowledge - Level of knowledge of the method in focus - Beliefs (including epistemological) - Gender - Ethnicity
Differences in context	<ul style="list-style-type: none"> - Size of student groups - Composition of student groups - Physical classroom context, access to artefacts/materials - Organisational structures - Degree and impact of external control/accountability measures, prevailing test discourse
Differences in content	<ul style="list-style-type: none"> - School subject or area in which the method is applied - Quality of the design of customised (externally created) teaching programmes and artefacts (including nature/length of intervention)

between students (e.g. reading or writing skills, dispositions, previous school success), instructional arrangements (e.g. number of students, material resources in the classroom), and the capabilities of those implementing instruction (e.g. beliefs about teaching and learning, success in managing the classroom, and experience from teaching writing and reading). (p. 737)

In many reviews, especially those of the past decade, research on the use of technological artefacts in instruction has been synthesised. All these reviews come to conclusions like that of Smetana and Bell (2012):

Despite the promise that computer simulations have shown in the teaching and learning of science, success is certainly not guaranteed. Like any other instructional resource, computer simulations can be effective if they are of high quality and are used appropriately. Therefore, the appropriate question for researchers is often how teachers and students use simulations, rather than whether the simulation in itself can achieve desired results. (p. 1362)

Overview finding 2: the need for highly qualified teachers

Overview finding 2 is linked to finding 1 and concerns the fact that moderating differences at the student level need to be recognised and compensated for by the teacher organising the instructional activities. In some of the underlying reviews, this is explicitly discussed (see [Appendix C1](#) “Differences in teachers”), but it is also a conclusion we draw on the basis of overview finding 1; since the effect of different methods is undoubtedly moderated by differences at the student level, the teacher’s ability to adapt and balance the use of a particular method is crucial. The aspects listed at the teacher level as competencies ([Table 1](#)) appear as central. Based on our analysis of the entire empirical material, we argue that there is high consensus (in terms of coherence) that no teaching method or artefact can replace a teacher who understands (1) that teaching (and hence the use of methods and artefacts) needs to be differentiated, and (2) that teaching not only involves conveying a given subject content according to a certain method or by using a certain artefact but also involves actively working to provide students with strategies for learning the content according to a method or artefact. The latter would also concern a gradual and conscious building of students’ metacognitive abilities.

As regards the teaching methods’ effectiveness in terms of students’ learning and development, our analysis shows that, above all, there is a distinction between students positioned as low-performing or diagnosed with some form of learning disability and students who are not so positioned or diagnosed. This is evident not least in reviews dealing with the relationship between direct and indirect instruction. Teaching methods based on constructivist thinking such as problem-based or discovery-based learning are, in our material, generally found to be less effective for lower-performing students or students diagnosed with different types of learning disabilities (e.g. Alfieri, Brooks, Aldrich, & Tenenbaum, 2011; Hmelo-Silver, 2004; Kirschner, Sweller, & Clark, 2006). It is also evident in several of the reviews concerning technology-related phenomena such as educational virtual reality, augmented reality, and computer simulation that there are several potential affordances in these for many students, while at the same time there are potential risks for others. Wu, Lee, Chang, and Liang (2013), for instance, discuss the crucial importance of teachers’ responsiveness to pupils’ different needs

when it comes to the use of technological artefacts in teaching (in this case augmented reality, or AR):

In an AR learning environment, students could be cognitively overloaded by the large amount of information they encounter, the multiple technological devices they are required to use, and the complex tasks they have to accomplish. That is, students need to be multi-tasking in AR environments. Dunleavy et al. (2009) reported that students often felt overwhelmed and confused when they were engaged in a multi-user AR simulation because they had to deal with unfamiliar technologies as well as complex tasks. Additionally, the tasks in AR environments may require students to apply and synthesize multiple complex skills in spatial navigation, collaboration, problem solving, technology manipulation, and mathematical estimation (Dunleavy et al., 2009). Previous research indicated that one reason for students’ learning challenges in AR environments lies in a lack of these essential skills (Kerawalla et al., 2006; Klopfer & Squire, 2008; Squire & Jan, 2007). Particularly for younger learners and novices at conducting open-ended investigations, additional scaffolding and support would be necessary to help them generate an appropriate plan of action, search for possible solutions to their problem, and interpret clues provided by the technological devices and embedded in the real-world environment (Klopfer & Squire, 2008). (p. 47)

The importance of the teacher is also underlined by Smetana and Bell (2012):

Even when support is provided by the simulation software and its accompanying materials, the teacher is critical for the successful implementation of instructional technologies and computer simulations in particular. There are no teacher-proof simulations. The teacher plays an important role in aligning the use of computer simulations to curricular objectives and to student needs. (Smetana & Bell, 2012, p. 1359)

Overview finding 3: the research–practice gap

A third theme cutting across many of the included research reviews concerns the research–practice gap. Despite the caveats discussed above and the impossibility of providing “correct” answers that apply once and for all, it can still be argued that research, over time, points to the relative benefits of some methods of teaching. However, the methods are not always put into practice by teachers, and the research–practice gap is problematised and discussed in several of the included reviews. Our analysis shows that the causes of the gap can be related to three (often interrelated) aspects (Table 2).

Table 2. Causes of the research–practice gap.

Research	<ul style="list-style-type: none"> - Insufficient didactic focus - Research not carried out in authentic environments - Insufficiently specific (need for greater specificity and/or consistency; need to consider more interacting moderators)
Teachers	<ul style="list-style-type: none"> - Insufficient knowledge - Insufficient or inadequate training - Strongly characterised by certain traditions or beliefs - Prefer to stay in the comfort-zone
Context	<ul style="list-style-type: none"> - System level: Strong external control, prevailing test discourse - System level: Insufficient knowledge at decision-making/policy levels - System level and school level: Inadequate organisational structures at schools - System level and school level: Inadequate physical and material conditions

Research is insufficiently didactic (a problem on the level of underlying studies)

A number of reviews problematise the fact that research on teaching methods is not carried out in actual classrooms. In fact, many of the underlying empirical studies included in the reviews that we have explored are conducted in laboratories, where the contextual conditions in many respects differ from those of everyday classrooms. Duit and Treagust (2003), for instance, conclude the following in their review on conceptual change in science education:

Educational research in general appears to be in danger of being viewed as irrelevant by many teachers (Lijnse, 2000). Kennedy (1997), for instance, argued that the 'awful reputation of educational research' (Kaestle, 1993) is due to the domination of basic research by cognitive psychology. Such studies are usually carried out in laboratory settings in order to allow strict control of variables. The price to be paid for a large degree of experimental 'cleanness' is that the results often do not inform the actual practice of teaching and learning. Wright (1993) provided similar arguments to explain that science education research is frequently viewed as irrelevant by policy makers, curriculum developers, and science teachers. He also claims that most science education researchers have little interest in putting into practice what is known. (p 681–682)

Rutten (2012), who reviewed the use of computer simulation in science education, argues:

The effects of computer simulations in science education are caused by interplay between the simulation, the nature of the content, the student and the teacher. A point of interest for the research agenda in this area, as mentioned by De Jong and van Joolingen (1998) in their review, is to investigate the place of computer simulations in the curriculum. Most of the studies we reviewed however, investigated the effects of computer simulations on learning *ceteris paribus*, consequently ignoring the influence of the teacher, the curriculum, and other such pedagogical factors. (p.151)

Research is insufficiently specific (a problem in underlying studies and on the review level)

The goal of most of the included 75 reviews is to be able to say something about the overall effect of teaching methods that are comprehensive by nature. *Scaffolding*, *problem-based learning*, *cooperative learning*, and *educational virtual reality* are some examples of such comprehensive methods. van de Pol, Volman, and Beishuizen (2010) reviewed a decades' research on scaffolding and conclude that scaffolding has neither been properly defined nor studied in a consistent way:

[A] challenge lies in documenting the effectiveness of the use of specific scaffolding strategies under particular circumstances empirically: Which strategies appear to work with which children in which grades and for which subject areas? (p. 287)

Among the studies included in each review, the composition of the overall studied population can range from pre-school children to adult students in higher education in different disciplines. Local contexts vary (sometimes strongly), due in part to the heterogeneity of the population but also due to a range of other factors. Additionally, the content of the studied interventions varies because of the methods' comprehensiveness. Formative feedback, as an example, can be given in a variety of ways (verbal, written, modelling, etc.); it can be provided from teacher to student, between students, or from computer to student. The extent of the feedback given can vary from

computerised, automated indications of the correctness of an answer to a factual question to lengthy written comments on an essay. The context in which feedback is given and received can also vary with regard to students' ages, school subjects, etc. The concrete and specific answers and guidelines that these types of studies can give to teachers battling with the how-questions of classroom practice are few. This fact is discussed by, for instance, Shute (2008), who concludes:

In general, and as suggested by Schwartz and White (2000) cited earlier, we need to continue taking a *multidimensional view of feedback* where situational and individual characteristics of the instructional context and learner are considered along with the nature and quality of a feedback message. (p. 176)

Young et al. (2012), who undertook a review based on the question of how effective video games are in enhancing students' learning, conclude by directing criticism to both themselves and the research community, urging researchers to "stop seeking simple answers to the wrong questions" (p. 83):

Video games vary widely in their design and related educational affordances: Some have elaborate and engaging backstories, some require problem solving to complete 5 to 40 multiplayer quests, and some rely heavily on fine motor controller skills. With this range of attributes, perhaps no single experimental manipulation (independent variable) can ever be defined to encompass the concept of video games writ large. Furthermore, given the diversity of student learning goals and abilities, likewise perhaps no singular outcome (dependent variable) from video games should be anticipated. Instead, applying principles from situated cognition suggests that research should focus on the complex interaction of player–game–context and ask the question, "How does a particular video game being used by a particular student in the context of a particular course curriculum affect the learning process as well as the products of school (such as test grades, course selection, retention, and interest)?" No research of this type was identified in our review, suggesting the missing element may be a more sophisticated approach to understanding learning and game play in the rich contexts of home and school learning. (p. 83–84)

Teachers and context

While many review authors agree that research has its clear part in the responsibility for the research–practice gap, it is also frequently suggested that lack of knowledge and training as well as lingering old traditions and beliefs among teachers are probable causes of the gap. Another aspect often discussed is how contextual factors affect teachers' teaching in a direction that neither research nor teachers themselves generally consider desirable. This is exemplified in the following excerpt from Driver, Newton, and Osborne (2000):

It is apparent that current classroom practices give little opportunity for young people to develop their ability to construct arguments. ... The major barrier to developing young people's skills of argument in science is the lack of opportunity offered for such activities within current pedagogical practices. If students are to be given greater opportunities to develop these skills, then this will require a radical change in the way science lessons are structured and conducted ... The fact that this does not happen in science lessons in the UK at the moment may be partly a reflection of the pressure that science teachers are under to "cover the National Curriculum." ... We suspect, however, that although these may be contributory factors, the main reason lies in the limitations of teachers' pedagogical repertoire and their limited understanding of the nature of science. Our observations and interviews with science teachers suggest that few teachers have the necessary skills to effectively organize group and class discussions and, hence, they lack confidence in their

ability to successfully manage sessions devoted to argumentation and discussion in the classroom. Consequently, such activities rarely, if ever, take place. (p. 308–309)

The recurrence over time

Finally, our analysis shows that similar issues/problems are identified in the conclusion and/or implication parts of the reviews *over time*. In the studies listed in [Appendix C2](#), the reader can note – especially as regards the role of research as the cause of the research-practice gap – that studies from all four decades are listed. Many conclusions concern the degree of classroom orientation in empirical research, and how great the opportunities are for research that is not conducted in actual classrooms to inform teachers' work in complex practice. To exemplify; in a review of instruction on problem-solving from the 1980s, Frederiksen (1984) concludes that researchers with different interests and approaches work in isolation from each other, and indeed in isolation from teachers and actual classroom settings, although it is precisely the collaboration between them that would be needed to provide answers that are more relevant to classroom practice. Similarly, Dole et al conclude in a review of reading comprehension instruction from the 1990s, that future research needs to be more classroom-oriented, didactic and specific. A clear example from the 2000s, is Duit and Treagust's claim that "[t]he price to be paid for a large degree of experimental 'cleanness' is that the results often do not inform the actual practice of teaching and learning" (2003, p. 681–682). And, finally, in a review of the use of computer simulations from the 2010s, Rutten (2012) refers to a previous review by de Jong and van Joolingen from 1998 where it was concluded that future research ought to investigate the place of computer simulations in the curriculum. 14 years later, in his own review, Rutten concludes that most studies still attempt to investigate the effects of computer simulations *ceteris paribus*, "consequently ignoring the influence of the teacher, the curriculum, and other such pedagogical factors" (p. 151). The above mentioned are examples of the fact that largely similar issues (which in many ways concern how the gap between theory and practice can be addressed by making primary research more classroom-oriented and specific) are discussed in research reviews over four decades.

Discussion

The purpose of the present study was to identify how issues related to the tension between contextuality and generalisability are elaborated in research reviews on teaching methods. Through careful mapping of the manifest data material, we have been able to show that such issues are frequently addressed and problematised in the analysed reviews. Three overview findings have been presented: the abundance of moderating factors, the need for highly qualified teachers, and the research-practice gap. In this final section, we will elaborate on our overview findings and discuss some implications for primary and secondary level research.

The abundance of moderating factors and the need for highly qualified teachers

Claiming that several factors affect the relationship between a teaching method and student learning is not very controversial. Methodologically, intervention

studies deal with a moderator as a third variable affecting the causal relationship between treatment (teaching method) and treatment outcome (effect on student learning). The fact that moderators are controlled for is in itself a recognition of the potential impact of the context. However, there is a difference between accounting for controlled moderators and explicitly problematising them in terms of what they may mean for a study's external and ecological validity.

We identified nearly 30 moderators addressed across the four areas of pupil, teacher, content, and context. Each of the moderators listed in [Table 1](#) is highly complex, and the number of possible combinations almost infinite. Obviously, it is difficult, not to say impossible, to determine with certainty the effect of a teaching method *ceteris paribus*. Simply put, methods do not have the same effect for all students in all situations. While this fact likely is self-evident to most (not least teachers), it seems necessary to repeatedly emphasise it in an era where the question asked often seems to be *What works?* rather than *What works for whom and in what circumstances?* (cf. Cartwright & Hardie, 2012; Pawson et al., 2005). Contextual variation and impact need to be clarified and acknowledged. In light of such recognition, a teacher can examine his/her own practice in relation to research findings and try to explore what will happen when employing a specific teaching method in his/her own context.

In the field of social work, Cartwright and Hardie (2017) propose a model aiming to predict the effect of a certain way of acting in a specific case. Such predictions, they argue, will require practitioners to draw heavily on their professional experience, causal understanding of their own situation, the proposed intervention, and its effects. More informed predictions may be made when intervention studies more fully account for the contextual complexity and circumstances. In a similar vein, Khorsan and Crawford (2014) discuss the importance of experimental studies in health care being explicit in explaining such aspects of the studies that are crucial for practitioners (as well as for secondary level researchers) if they are to be able to judge the external validity of implementation and outcomes. They argue that study quality must be regarded as a multidimensional concept that includes both internal, external (population) and ecological (situation and setting) validity. Moreover, they propose an external validity assessment tool to measure the extent to which and how well various context and intervention characteristics are described in experimental studies. Only if such aspects are clearly described, the judgement of relevance for other settings is made possible. In the field of teaching and learning, Bernstein (2018) discusses generalisation as a two-way street, where the possibility to judge the external validity of a study is a shared responsibility between the author and the reader of a study. The author's responsibility is to provide enough information in terms of rich, thick descriptions of context to make judgements about generalisation possible. However, the responsibility for discerning useful parts of the study and relating them to other contexts rests with the reader.

In the section of overview findings, we argue that no teaching method or artefact can replace the context-experienced teacher. The effect of methods on students' learning is undoubtedly moderated by differences at the student level and other factors, wherefore the teacher's situational awareness and ability to predict or know what may work for whom, how, and in what circumstances is crucial. The

teacher definitely needs the method, and the method certainly needs the reflective teacher. In line with the arguments above from researchers in different fields, we find it important not only to account for moderating factors, but also to explain and problematise the complexity of the context in such a way that practitioners within the field of teaching may assess the external and ecological validity of a study.

The research-practice gap

The identification of a gap between research and practice is neither new, nor unique to the research field of teaching methods. We have shown that the gap is attributed on the one hand to factors at the levels of the teacher and the context, and on the other hand to the fact that primary research is insufficiently naturalistic, didactic and specific. The high degree of experimental cleanness that can be achieved in, for example, the laboratory environment may ensure high internal validity, but simultaneously limit the external or ecological validity of the findings (Bernstein, 2018; Khorsan & Crawford, 2014). The failure of primary research to validate why a particular teaching method worked or not in a certain context is emphasised in many reviews as a cause of the research-practice gap. In addition, some reviews contain elements of self-criticism against the secondary level of research.

We too would like to suggest that the long-standing gap between research and practice is an issue that needs to be addressed in both first- and second-order research. In the introduction, we argued that producing reviews is a logical and reasonable way to integrate findings and insights from different studies, and that systematic research reviews can contribute in various ways with knowledge that can potentially bring research forward and inform both practice and policy. According to Gough et al. (2012, p. 5), research reviews are vital for various reasons:

[R]eviews enable us to establish not only what is known from research; but also what is not known. They can inform decisions about what further research might be best undertaken, thereby creating a virtuous cycle. They enable researchers, policymakers and practitioners to answer key questions: ‘what do we know, how do we know it?’ and ‘what more do we want to know and how can we know it?’

Reviews are crucial for establishing what is known and not known. The reviews we have analysed are often adequately cautious in their conclusions of what is known. Moreover, they point out knowledge gaps and how these can or should be addressed in future research. However, by analysing a sample of research reviews spanning a period of four decades, it becomes clear that the same types of problems and knowledge gaps are pointed out repeatedly. The virtuous circle mentioned by Gough et al. (2012) above is a metaphor used to underscore that one does not arrive at the same point, but rather that there is a continuous knowledge development. The idea of research reviews as an important element in creating virtuous circles presupposes that drawn conclusions and appeals made in reviews form (at least in part) the starting point for new primary studies. Reasonably, the primary study level has a great deal of responsibility when it comes to creating more context-specific knowledge about teaching methods. However, the responsibility also lies with second-order research and how the tension between contextuality and generalisation is handled there.

A contextually bound approach to causality

Pawson (2006) and Pawson et al. (2005) argue that few clues are given in traditional research reviews as to why interventions in the health service field show different and sometimes even contradictory results in different contexts. According to the authors, the quest to understanding “what works” needs to be more realistic and they advocate a realist inquiry model in reviews. The model takes its starting point in a different understanding of causality than the traditional review approach. Transferred to the context of teaching methods, one needs to understand the underlying mechanisms that connect the teaching method and its effect on a student’s learning, as well as the context of the relationship. These interrelationships must be hypothesised and examined. Consequently, the question of “what works?” changes to “what works for whom in what circumstances?” Further, Pawson and colleagues argue that the conceptualisation of interventions as dynamic and complex systems-within-systems imposes certain limitations or requirements on a reviewer. First, there is a limit to how much territory can be covered in a review. It must be clearly articulated in the review question(s) precisely which aspects of an intervention or method are being studied. Second, reviews cannot focus solely on tangible processes and easily measured outcomes but need also to concentrate upon more subtle contextual conditions. Finally, the reviewer must be cautious when delivering recommendations. The realist review approach that Pawson and colleagues advocate delivers “illumination rather than generalizable truths and contextual fine-tuning rather than standardization”, which brings with it that “[h]ard and fast truths about what works must be discarded in favour of contextual advice in the general format: in circumstances such as A, try B, or when implementing C, watch out for D” (p. 24). Thus, a realistic review focuses on explaining contextual complexity in such a way that it allows the reader to make more informed choices (see also Rycroft-Malone et al., 2012).

Another prominent issue pointed out in the findings of our study which needs to be considered at both primary and secondary levels, is the need for greater specificity. In the section of overview findings, we argue (using the example of “formative feedback”) that many of the teaching methods are comprehensive by nature and include many different aspects. Additionally, there is sometimes a lack of consensus regarding the definition and demarcation of a given method. It becomes highly problematic when various studies that state the effect of a method (or, as it happens, even argue for or against its existence) may not refer to or have studied the exact same thing. These problems are highlighted fairly consistently by the authors in the reviews included in our data. The fact that it is problematic with blurry definitions becomes especially clear regarding student-centred teaching methods such as inquiry-based learning, discovery-based learning and problem-based learning (e.g. Alfieri et al., 2011; Furtak, Seidel, Iverson, & Briggs, 2012; Hmelo-Silver, 2004; Kirschner, 2006; Minner, Levy, & Century, 2010).

To conclude

Through our overview findings, we have highlighted issues that are frequently problematised across high impact research reviews on teaching methods over a period of four

decades. The substantive aspects of the findings are neither surprising nor previously unknown. The strength of this study lies in how we have been able to show patterns and coherence in conclusions *across* studied issues over time and their relevance for the tension between context and generalisation.

Trying to determine where the effect of a method itself ends and where the impact of the context begins is perhaps a mission impossible. What can be done in both primary and second-order research is to explicitly recognise (to a greater extent), explore, and discuss contextual complexity. In line with other researchers above referred to, we want to underline the importance of viewing validity as a multidimensional concept including both internal, external, and ecological aspects. Basically, there are two questions research on teaching methods ought to respond to: whether a particular way of teaching has an impact on students' learning and performances, and what and how others can learn from completed studies. Both are equally important, but the internal validity of studies seems to be more valued than the external and ecological validity. As Bernstein (2018) argues, foregrounding one at the expense of the other does not help advancing the field of knowledge:

If we are unable to determine if what we are doing is working, we exist in an evidence-free zone in which we are grasping in the dark to find the most effective ways to teach our content. In addition, if we are unable to generalize our work to other contexts, we are not building a field, and are not allowing the practice of teaching to advance outside our individual classrooms. (p. 123)

Thus, richer descriptions and problematisation of context are needed, for both practitioners and reviewers to be able to determine validity in a multidimensional way. As for the review level, the realist approach suggested by Pawson et al. (2005) may well be a viable way forward also in the field of research on teaching methods. Not least – and due to the fact that many teaching methods are both comprehensive and complex – it is important to emphasise the need for clearly articulated research questions stating which aspects of an intervention or method are being studied and that there is a limit to how much territory a review can cover.

Funding

This work was supported by the Vetenskapsrådet [2016-03679].

Notes

1. We are well aware of the fact that the WoS covers far from all educational research; nevertheless, we restricted our searches to it because of its acknowledged high quality and its prestigious position among databases.
2. A number of included reviews are based on studies carried out in both K-12 context and in higher and/or adult education.
3. The codes are basically those listed as dashes under the four subcategories of overview finding 1 (Table 1) and under the three subcategories of overview finding 3 (Table 2) in the results section.
4. The CERQual framework primarily concerns reviews (i.e. secondary level) and involves assessment of the methodological limitations and adequacy of data in underlying empirical

qualitative studies. This has not been relevant in our case; our use of the CERQual's starting points concerns the coherence of the overview findings.

5. Underlying studies (in their entirety) cannot in our case be described as “qualitative research”. However, the data we have analysed are qualitative (i.e. text excerpts).
6. Since our three overview findings are to a certain extent linked to each other, the reader will notice that some of the excerpts in the results section are in fact illustrative of more than one overview finding.

Notes on contributors

Åsa Hirsh is Associate Professor in Education at Jönköping University and the University of Gothenburg, Sweden. Her research focus lies at the intersection between the research fields of classroom instruction, educational assessment, and school development.

Claes Nilholm is a professor of Education at Uppsala University, Sweden. His research focus is on inclusive education. He also has an interest in methodological issues in research reviewing and is currently leading the project “Research about teaching – systematic mapping and analysis of research topographies” financed by the Swedish Research Council, educational sciences.

Henrik Román is senior lecturer in Education at Uppsala University, Sweden. His research focuses historical aspects of contemporary educational policy and practice.

Eva Forsberg is professor in Education at Uppsala University, Sweden and general editor of Nordic Journal of Studies in Educational Policy (NordSTEP). Her research focuses the interface between educational policy, practice and research from a curriculum theoretical perspective.

Daniel Sundberg is Professor of Education, Linnaeus University, Department of Education and Teachers' Practice, Sweden. He is co-leader of the SITE research group (Studies in Curriculum, Teaching and Evaluation) and chief editor of Educational Research in Sweden. His main field of research is comparative and historical perspectives on education reforms, curriculum and pedagogy.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Claes Nilholm  <http://orcid.org/0000-0001-8613-906X>

Eva Forsberg  <http://orcid.org/0000-0002-1768-1450>

Daniel Sundberg  <http://orcid.org/0000-0003-0644-3489>

References

- Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, 103(1), 1–18.
- Barber, M., & Mourshed, M. (2007). *How the world's best-performing school systems come out on top*. London: McKinsey & Co.
- Bernstein, J. L. (2018). Unifying SoTL methodology: Internal and external validity. *Teaching & Learning Inquiry*, 6(2), 115–126.
- Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. New York: Oxford University Press.

- Cartwright, N., & Hardie, J. (2017). Predicting what will happen when you intervene. *Clinical Social Work Journal*, 45(1), 270–279.
- Cavanagh, S. (1997). Content analysis: Concepts, methods and applications. *Nurse Researcher*, 4(3), 5–16.
- Cobb, B., Lehmann, J., Newman-Gonchar, R., & Alwell, M. (2009). Self-determination for students with disabilities: A narrative meta-synthesis. *Career Development of Exceptional Individuals*, 32(2), 108–114.
- Coffey, A., & Atkinson, P. (1996). *Making sense of qualitative data. Complementary research strategies*. Thousand Oaks, London, New Delhi: Sage Publications Inc.
- De Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, 68(2), 179–201.
- Dole, J. A., Duffy, G. G., & Pearson, P. D. (1991). Moving from the old to the new: Research on reading comprehension instruction. *Review of Educational Research*, 61(2), 239–264.
- Downe-Wamboldt, B. (1992). Content analysis: Method, applications, and issues. *Health Care for Women International*, 13(3), 313–321.
- Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, 84(3), 287–312.
- Duit, R., & Treagust, D. F. (2003). Conceptual change: A powerful framework for improving science teaching and learning. *International Journal of Science Education*, 25(6), 671–688.
- Dunleavy, M., Dede, C., & Mitchell, R. (2009). Affordances and limitations of immersive participatory augmented reality simulations for teaching and learning. *Journal of Science Education and Technology*, 18(1), 7–22. <http://dx.doi.org/10.1007/s10956-008-9119-1>
- Frederiksen, N. (1984). Implications for cognitive theory for instruction in problem-solving. *Review of Educational Research*, 54(3), 363–407.
- Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research*, 82(3), 300–329.
- Gough, D., Thomas, J., & Oliver, S. (2012). Clarifying differences between review designs and methods. *Systematic Reviews*, 28(1), 1–9.
- Graham, S., & Hebert, M. (2011). Writing to read: A meta-analysis of the impact of writing and writing instruction on reading. *Harvard Educational Review*, 81(4), 710–744.
- Graneheim, U. H., & Lundman, B. (2004). Qualitative content analysis in nursing research: Concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today*, 24(2), 105–112.
- Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, 26(2), 91–108.
- Hargreaves, A., & Fullan, M. (2012). *Professional capital. Transforming teaching in every school*. New York: Teachers College Press.
- Hattie, J. (2003). Teachers make a difference: What is the research evidence? *Paper presented at the Building Teacher Quality: What does the research tell us ACER Research Conference*, Melbourne, Australia. Retrieved from http://research.acer.edu.au/research_conference_2003/4/
- Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.
- Hirsh, Å., & Nilholm, C. (2019). Reviews of teaching methods – what are the fundamental problems? *Paper presented at ECER Conference*, Hamburg. Retrieved from <https://eera-ecer.de/ecer-programmes/conference/24/contribution/47337/>
- Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review*, 16(3), 235–266.
- Hsieh, H., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288.

- Kaestle, C. F. (1993). The Awful Reputation of Education Research. *Educational Researcher*, 22(1), 23–31. doi:10.3102/0013189X022001023
- Kazrin, A., Durac, T., & Agteros, T. (1979). Meta-meta-analysis: A new method for evaluating therapy outcome. *Behaviour Research & Therapy*, 17(4), 397–399.
- Kennedy, M. M. (1997). The connection between research and practice. *Educational Researcher*, 26(7), 4–12. doi:10.3102/0013189X026007004
- Kerawalla, L., Luckin, R., Seljeflot, S., & Woolard, A. (2006). “Making it real”: exploring the potential of augmented reality for teaching primary school science. *Virtual Reality*, 10(3), 163–174. doi:10.1007/s10055-006-0036-4
- Khorsan, R., & Crawford, C. (2014). External validity and model validity: A conceptual approach for systematic review methodology. *Evidence-Based Complementary and Alternative Medicine*, 2014, 1–12.
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2), 75–86.
- Kondracki, N. L., Wellman, N. S., & Amundson, D. R. (2002). Content analysis: Review of methods and their applications in nutrition education. *Journal of Nutrition Education and Behavior*, 34, 224–230.
- Lewin, S., Glenton, C., Munthe-Kaas, H., Carlsen, B., Colvin, C. J., Gulmezoglu, M., ... Rashidian, A. (2015). Using qualitative evidence in decision making for health and social interventions: An approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). *PLoS Med*, 12(10), 1–18.
- LIJNSE, P. (2000) Didactics of science: The forgotten dimension in science education research. In Millar, R., Leach, J., & Osborne, J. (Eds.), *Improving science education* (pp. 308–326). Buckingham, UK: Open University Press.
- Maag, J. W. (2006). Social skills training for students with emotional and behavioral disorders: A review of reviews. *Behavioral Disorders*, 32(1), 5–17.
- Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction-what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching*, 47(4), 474–496.
- Nilholm, C., & Göransson, K. (2017). What is meant by inclusion? An analysis of European and North American journal articles with high impact. *European Journal of Special Needs Education*, 32(3), 437–451.
- OECD. (2016). *PISA 2015 results (Volume II): Policies and practices for successful schools*. Paris: Author. doi:10.1787/9789264267510-en
- Pawson, R. (2006). *Evidence-based policy*. London: SAGE Publications Ltd.
- Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2005). Realist review – A new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy*, 10(1), 21–34.
- Polanin, J., Maynard, B., & Dell, N. (2017). Overviews in education research: A systematic review and analysis. I. *Review of Educational Research*, 87(1), 172–203.
- Roman, H., Sundberg, D., Hirsh, Å., Nilholm, C., & Forsberg, E. (2018). Mapping and exploring high impact research reviews on teaching. *Paper presented at ECER Conference*, Bolzano. Retrieved from <https://eera-ecer.de/ecer-programmes/conference/23/contribution/44956/>
- Rutten, N., van Joolingen, W. R., & van der Veen, J. T. (2012). The learning effects of computer simulations in science education. *Computers & Education*, 58(1), 136–153.
- Rycroft-Malone, J., McCormack, B., Hutchinson, A. M., DeCorby, K., Bucknall, T. C., Kent, B., ... Wilson, V. (2012). Realist synthesis: Illustrating the method for implementation research. *Implementation Science*, 7(33), 1–10.
- Saini, M., & Shlonsky, A. (2012). *Systematic synthesis of qualitative research*. New York: Oxford University Press.
- Schwartz, F., & White, K. (2000). Making sense of it all: Giving and getting online course feedback. In White K. W. & Weight B. H. (Eds.), *The online teaching guide: A handbook of*

- attitudes, strategies, and techniques for the virtual classroom (pp. 57–72). Boston: Allyn & Bacon.
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189.
- Smetana, L. K., & Bell, R. L. (2012). Computer simulations to support science instruction and learning: A critical review of the literature. *International Journal of Science Education*, 34(9), 1337–1370.
- Squire, K., & Jan, M. (2007). Mad city mystery: developing scientific argumentation skills with a place-based augmented reality game on handheld computers. *Journal of Science Education and Technology*, 16(1), 5–29. <http://dx.doi.org/10.1007/s10956-006-9037-z>
- Stigler, J. W., & Hiebert, J. (2009). *The Teaching gap: Best ideas from the world's teachers for improving education in the classroom. Updated with a new preface and afterword*. New York: Free Press.
- Terhart, E. (2011). Has John Hattie really found the holy grail of research on teaching? An extended review of Visible Learning. *Journal of Curriculum Studies*, 43(3), 425–438.
- Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8(45), 1–10.
- Torgerson, C. J. (2007). The quality of systematic reviews of effectiveness in literacy learning in English: A 'tertiary' review. *Journal of Research in Reading*, 30(3), 287–315.
- van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of Research. *Educational Psychology Review*, 22(3), 271–296.
- Wright, E. (1993) The irrelevancy of science education research: perception or reality? *NARST News*, 35(1), 1–2.
- Wu, H. K., Lee, S. W. Y., Chang, H. Y., & Liang, J. C. (2013). Current status, opportunities and challenges of augmented reality in education. *Computers & Education*, 62, 41–49.
- Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., ... Yukhymenko, M. (2012). Our princess is in another castle: A review of trends in serious gaming for education. *Review of Educational Research*, 137160065182(1), 61–89.