Policy-based evidence? The schools inspectorate in England, research and school mathematics policy.

Ashley Compton[a]* and Mark Boylan[b]

[a]*Bishop Grosseteste University, Lincoln, UK;* [b]*Sheffield Institute of Education, Sheffield, UK*

Corresponding author

Mark Boylan

m.s.boylan@shu.ac.uk

Sheffield Hallam University

Arundel 10107

Sheffield

S1 1WB

Policy-based evidence: The schools inspectorate in England, research and school mathematics policy.

Increasingly, debates about school mathematics curriculum and pedagogy reference evidence and research. Policymakers and others seek to influence practice is through regulatory bodies such as inspection services, like the Office for Standards in Education (Ofsted) in England. In 2021, Ofsted published a series of curriculum research reviews, including one focused on mathematics. Using features of education review quality, we analyse Ofsted's departure from accepted ways to synthesise research in the creation of policy-based evidence. Ofsted disregards usual scholarly norms of research with a lack of transparency, weakness of research design and search strategy, and lack of rigour in the selection of evidence. Further, much of the cited research is misinterpreted and misused, and unwarranted causal claims are made with overgeneralisation and oversimplification being a consistent thread through the review. A specific exemplification of this is the misappropriation of research on problem solving leading to recommendations conflicting with both the National Curriculum in England and the findings of other, more rigorous, research reviews. From this analysis, we argue that the Mathematics Review is an example of 'policy-based evidence' and point to ways that inspection evidence can complement mathematics education research to support evidence informed policy and practice.

Keywords: evidence, policy, inspection

Introduction

Mathematics education internationally is influenced by curriculum policy, discourses of evidence-based practice and school inspectorates. Our focus in this paper is an example where these three influences intersect: a research review undertaken by The Office for Standards in Education (Ofsted), the school inspection service in England. To inform inspections and shape practice, Ofsted undertook a series of subject research reviews including one for mathematics (Ofsted, 2021a), which aimed to synthesise research into factors affecting the quality of mathematics education. We examine how justifiable Ofsted's claims are in the mathematics review and in so doing we analyse an example of the politicisation of evidence in education policy.

Education systems internationally vary to the extent that they are politically directed, the degree of autonomy schools and teachers have, and levels of accountability. The education system in England has been characterised as having high autonomy with high accountability (Malin, et al., 2020). High autonomy because the majority of schools have become part of publicly funded but independent trusts, which often comprise of multiple schools (Benn, 2023). Trusts have considerable apparent freedom over curriculum and pedagogy, although individual schools have varying degrees of autonomy within trusts. High accountability because of a high-stakes system of inspection and performance measures, with Ofsted having considerable power and influence over schooling (Perryman et al., 2018). The accountability systems are important to government in England exercising considerable political and ideological direction of the curriculum and teaching approaches, in spite of rhetoric of school autonomy.

Ofsted's influence on schools is most explicitly exercised through its grading of schools and educational settings, which is important to their status and can lead to changes in school-governance (McVeigh, 2020). The importance of inspection services

in influencing education is not unique to England (Gustaffsson et al. 2015; Ozga, 2014) and like inspection systems internationally, Ofsted also shapes education through setting expectations (Gustaffsson et al., 2015). A key regulatory tool is the frequently changed Inspection Framework, comprising the process and criteria for judging schools (McVeigh, 2020); the inspection framework of 2019 (Ofsted, 2019a) sought to more directly influence teaching content with a focus on an advocated approach to curriculum and teaching methods and through its mobilisation of the warrant of evidence in claims to be research-based (Spielman, 2023).

Internationally, discourses of 'evidence' and 'evidence-based' have become prominent in education policy (Helgetun & Menter, 2022). In education, in England, this is apparent through the prominent role of the Education Endowment Foundation (EEF). The partially government-funded EEF aims to generate evidence through Randomised Control Trials and systematic reviews, as well as supporting evidence mobilisation (Coldwell, 2022). The EEF is one of a network of policy driven 'What Works Centres' in the UK and is part of an international network of similar organisations, including at least 12 What Works Clearinghouses engaged with educational practice in the USA (Wadhwa et al, 2024).

A recent policy preference for 'traditional teaching' methods such as explicit instruction is justified by appeals to research and claims that these practices are evidence-based (Gibb, 2015) and is also advocated by other influential groups and networks (Benn, 2023). This is most obvious in relation to reforms of initial teacher education (Craske, 2021; Hordern & Brooks, 2023; Turvey, 2023). Similar movements are found elsewhere, for example in the USA (The Science of Math, n.d.). Against this background, 'evidence' and 'evidence-based' have multiple and contested meanings. This is not unique to education. Indeed, Strassheim & Kettunen (2014) identified ways

that evidence is produced to justify or mobilise for a policy position, which they call policy-based evidence in contrast to the more usual evidence-based policy. Our central argument in this paper is that the Ofsted review is policy-based evidence.

Generally, mathematics educators, subject associations and independent organisations that provide professional development in England responded negatively to Ofsted's recommendations (for example, Early Childhood Maths Group, 2021; Tyler, 2021; Watson, A., 2021; Watson & Back, 2021). Alongside the concerns about Ofsted's recommendations for mathematics teaching, issues were identified about the quality of scholarship, such as including poor-quality studies and generalising from these (Gilmore et al., 2021). Members of the Association of Mathematics Education Teachers (AMET) identified several instances of questionable use and interpretation of sources in the review, which prompted a detailed analysis involving both authors. (Compton, 2021).

In this paper, we develop these concerns by analysing ways Ofsted used evidence and how this compares to customary ways to synthesise research and infer implications for practice. We argue that the Ofsted review is an example of policy-based evidence and propose ways for a more productive use of research by inspectorates. Our analysis below focuses on the purpose and methodology to the review rather than on individual researchers tasked with conducting it. Although we are critical of the review practices, we also acknowledge a positive feature of the review is that it did seek to embrace a wide range of sources of evidence including qualitative, methodologies.

Inspection, knowledge and Ofsted's mathematics review

As noted, inspection bodies influence schools by setting expectations (Gustaffson, 2015), and in England this is through the inspection framework. However, inspection

bodies internationally implement public policy through epistemic activity in other ways. As well as knowledge produced for and from regulation, inspection services may have inquisitorial and meditative functions (Baxter and Hult, 2017; Jacobsson, 2006; Ozga, 2014). Inspection's regulative purpose is its most visible and central power in school systems. Important too is inquisitorial activity. This generates evidence and knowledge at a system level by auditing (Jacobsson, 2006) school practices through inspection, a practice found across European education systems (Baxter and Hult, 2017; Ozga, 2014). What Jacobsson (2006) refers to as meditative activities, consists of reflection on lessons learned from regulation and audit, through sharing of ideas and experience (Sasen, 2007). An example of this meditative activity is a report on current practice in mathematics education (Ofsted, 2023).

All these functions entail the production and communication of knowledge, valuable functions of inspection bodies. Additionally in England the inspectorate has adopted a role brokering knowledge produced by others, and importantly in the context of the Ofsted mathematics review, knowledge produced through academic research. Previously, Ofsted commissioned researchers to do this, such as a series of research reviews in the 1990's, including one in mathematics (Askew & Wiliam, 1995) and a comparative review of international practice (Reynolds, 1996). So, Ofsted seeking to influence school practice through evidence reviews is not itself a departure from previous practice (Smith, 2000), but how this was done through the 2021 to 2023 research reviews, as we later argue, was different.

The scope of the mathematics review was ambitious, with four main questions, and 30 more detailed related sub-questions. The main questions are shown in Figure 1, with an illustrative selection of sub-questions that refer to problem solving, exemplified

in our analysis. Note that these questions were not published in the review but were provided following a Freedom of Information (FOI) request.

Figure 1. Ofsted review questions and selection of sub-questions

Main questions

1. What is it that explains the variation in the quality of mathematics education in schools? (2 sub-questions)

2. How does variation in understanding of curriculum progression in this subject affect the quality of mathematics education? (8 sub-questions)

3. What are the main pedagogical questions in this subject, and what impact do those dimensions have on variations in quality? (5 sub-questions)

4. How can approaches to assessment (7 sub-questions), systems at subject level and whole school policy decisions affect quality of education in mathematics? (8 sub-questions)

Selection of sub-questions

2c What are the best approaches to choosing and sequencing content that enables reasoning and problem solving?

3c Which pedagogical approaches are most effective at helping pupils to reason and solve problems?

4c Which summative assessment approaches are most effective for ascertaining whether pupils can reason and solve problems?

Reviewing the review: methodology

We reviewed the Ofsted mathematics report to assess the extent to which claims made in the review were justifiable and investigated the review as a policy text. We used two principal methods: analysing review quality and policy text analysis. Although they are

presented here sequentially, they were undertaken in parallel and influenced each other. The analysis of review quality consisted of identifying features of review quality, citation analysis and reproducing searches based on the response to the FOI request to Ofsted.

1. Analysis of review quality

Features of review quality

The Ofsted subject review differs from academic reviews of research studies such as systematic reviews of the literature or indeed literature reviews in a traditional sense. The National Director, Education for Ofsted stated, "the term 'research review' is not generally used in any fixed way and it is simply descriptive in this context" (Harford, 2021, p.2). In a Twitter exchange about whether these documents were research reviews, the leader of Ofsted's Curriculum Unit, said "We did not want people to think they were literature reviews in the true sense hence not calling them that. I'd agree they are something like a position paper" (Fearn, 2021).

However, naming the text a 'research review', producing a principles document outlining bodies of literature, using search strategies and the textual devices of citation indicate that this was meant to be read as scholarly text. Further, the overall framing of Ofsted's inspection approach was research-based (Spielman, 2023). Considering Ofsted's disavowal that it is a traditional literature review, so excluding that 'review family', the review may be considered 'a purpose specific review' (Sutton et al., 2019).

There are protocols for systematic reviews and meta-analyses, such as Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA, 2024). However, for other types of review that may influence policy such standards are not yet established. The original AMET review, that both authors were involved in, drew on the

team's academic experience as researchers and university lecturers in analysing Ofsted's document. However, this approach has been further developed during the production of this article, drawing upon several sources. We drew on guidance for method and quality for synthesis in educational reviews (Alexander, 2020; Murphy et al., 2020), including educational reviews adopting an inclusive approach to the type of methods leading to the outputs reviewed (Suri & Clarke, 2000). We considered the National Research Council's (2001) mathematics education review that included a guidance section for others, in addition to substantive review findings and compared it to the Ofsted review. We also considered analyses of citation practices, and academic attribution (Hyland, 1999; West et al., 2017), as these were concerns raised in early responses to the Ofsted review. Using these sources we developed a framework of eight features in an iterative process from the original AMET review, further analysis for this article and feedback from reviewers to create a more neutral response. Table 1 lists these features and a summary of indicators with our sources. We integrate into our analysis below why these are important and evaluative assessments of the extent to which these features are present in the Ofsted review.

Table 1. Features of review quality

Features	Indicators
Transparency	Questions posed are clearly articulated (Alexander, 2020); methodology is explicit including how the literature was identified (Alexander, 2020; Murphy et al., 2020)
Review design	Design and review questions are aligned (Alexander, 2020)
Search strategy	Inclusion of search terms, (Murphy et al., 2020) Use of multiple databases for inclusivity and complement with journal searches (Alexander, 2020)

Selection of sources	Appropriate inclusion and exclusion criteria, considering aspects such as output age and quality, with older sources potentially not relevant given educational changes (Alexander, 2020)
Citation practices	Accurate crediting of researchers' contributions (West et al., 2017) Disciplinary norms are followed (Hyland, et al., 2009)
Interpretation of research	Researchers whose work is reviewed are likely to agree with interpretation of their research (Suri & Clarke, 2000) Drawing conclusions based in the literature (Alexander, 2020)
Causal claims	Claims should be made according to the methods and quality of reviewed research (NRC, 2001; Suri & Clarke, 2009)
Generalisation	Describe limitations of applicability of statements (NRC, 2001; Suri & Clarke, 2009) Consider applicability to context of potential users, particularly normal classroom practices (NRC, 2001)

Citation analysis

The AMET team read, and coded, 235 sources related to the 307 citations within the 201 footnotes of the review. The following codes were used:

- source characteristics (type of text, e.g. research-based article or policy paper, date of publication, geographical origin, age phase, evidence base for claims).
- the degree of alignment between the Ofsted review claim and the cited source

Each text was analysed as to how far the contents of the source matched the claim in the Ofsted review related to that citation, with these classified as 'full', 'partial', or 'no match'. The differences between these categorisations are exemplified

later when the findings of the analysis are discussed. For research sources, the methodology and sample size were considered and whether these justified claims, including casual claims. This analysis was revisited and refined in the production of this paper.

Some of the authors whose work we determined had been misused were contacted to corroborate our interpretation of their work compared to that of Ofsted. Although this is not common practice in research reviews, it fits with Suri and Clarke's (2000) criteria that researchers should agree with the interpretation of their work. The selection here was somewhat opportunistic, dependent on our networks or those of other mathematics educators who shared similar concerns to ours about the review.

Review of databases used and replication of searches

The FOI request identified the data bases used and search strategy. We attempted to replicate the searches undertaken to gain insight into selection of texts included by Ofsted.

2. Policy text analysis

To understand the background to the review, and what had influenced both its production and the review as a text, we undertook an analysis of policy texts, both those produced by policy actors and responses to these (Boylan & Adams, 2023; Rizvi & Lingard, 2010). These included:

1) Published Ofsted reports and papers such as the research basis for the inspection framework, the principles underpinning the subject reviews, other subject reviews, previous Ofsted research);

2) More informal Ofsted texts such as a speech setting out Ofsted's positions on evidence, responses to complaints and social media commentary, and the FoI response;
3) Responses by mathematics educators to the review, supplemented by texts raising concerns about other Ofsted subject reviews;
4) Background policy texts such as the National Curriculum for England, ministerial speeches and policy announcements, and academic research and commentary on relevant educational policy.

Ofsted's review practices

In this section, we apply our framework and examine Ofsted's research practices in the mathematics review. Given the limits of this article, we provide examples of these issues rather than a full account of each, selecting those both illustrative and understandable as standalone examples.

1. Transparency

Transparency is an important aspect of research integrity (Alexander, 2020; Murphy et al, 2020; NRC, 2001; Suri & Clarke, 2009) and one that the British government is committed to (HCST, 2018). However, the Ofsted (2021a) review stated little about the review questions, methods, or selection criteria. Ofsted director Harford (2021) stated that they did not provide transparency about their deliberations regarding inclusion of literature because this would make the review more accessible for the reader. A subject review 'Principles' document (Ofsted, 2021b) provided general outlines of the approach across all the subject reviews but did not provide the specific mathematics research questions, the search terms or databases used, nor the inclusion and exclusion criteria. This contrasts with greater methodological transparency in other mathematics reviews,

such as Adding it up! (NRC, 2001), a Nuffield series of review papers (Nunes et al., 2008) and an EEF review (Hodgen et al., 2018). The NRC (2001) report was acknowledged as an expert judgement review rather than a research review. Nunes et al. (2008) included a full paper which set out the methodological approach, some inclusion and exclusion criteria and the sources searched. However, Nunes et al. (2008) could have been even more transparent had it included the search strings, dates, detailed inclusion / exclusion criteria and the complete database of papers. The EEF review explained the types of studies used, the scope of searches, details of methods and their limitations, and included a guide for teachers on how to interpret the report and the evidence presented. This was supported further by details about the search strategy in the appendices. Ofsted's lack of inclusion of research questions necessitated the FOI request.

2. *Review design*

According to Alexander (2020) it is important to devise questions that are answerable within the context of the review. The FOI request produced the 4 main questions (see Figure 1) and 30 more detailed sub-questions. The questions were both numerous and broad, making the scope of this review unwieldy. Each of the review questions is worthy of a focused review but in a 10,000-word document there were only ~300 words available per question. This suggests that the design was too broad. In a section on inclusion, 131 words are devoted to generalised claims recommending explicit instruction to support learners with Special Educational Needs and Disabilities. This is followed by 179 words focused on autism. However, notably absent is any discussion of dyscalculia or dyslexia.

For all the subject reviews, a Principles paper set out the overall approach explaining that the reviews would use research that "...aligns with the principles for

quality of education, as outlined in the education inspection framework" (Ofsted, 2021b, p.4). Alexander (2020) discussed the danger of authors' theoretical position influencing the design and search strategy to the extent that it becomes an argumentative review intended "to share their 'truth' with others" (p.10).

3. Search Strategy

The FOI request provided search strategy details which had been lacking in the review. Ofsted used a single database, Education Research Complete (ERC) (FOI request), plus sources from EEF, Nuffield Foundation, Ofsted, DfE, PISA and TIMSS (Ofsted, 2021b). The strategy contrasts with the EEF review (Hodgen et al., 2018) which used 15 data bases, in addition to some specific sources. ERC includes some peer-reviewed journals, such as Educational Studies in Mathematics (ESM) and Research in Mathematics Education but is missing other important journals, such as Journal of Research in Mathematics Education, ZDM and Journal of Mathematical Behaviour.

The use of only one database will have narrowed the potential scope of the search, although the additional sources listed would help. According to the FOI request, 35 searches were conducted in total, 34 of them using a single source. The other search used additional sources such as google scholar. Of the 34 using a single source, 26 used the ERC database. The other 8 searches used additional sources such as EEF's repository of guidance and Nuffield Foundation's archive, but also included a "collection of papers" (FOI), with the source and content unclear. The overall approach suggests that the review design was to curate sources to evidence existing positions.

It was not clear from reproducing some of these searches how the papers used were selected, especially as several ESM published papers identified in the reproduced searches were more aligned with Ofsted's positions but not included. The section on mathematics anxiety did not include a Nuffield published report (Carey et al., 2019) on

mathematics anxiety, although this would have appeared in the Nuffield Foundation archive search. There are numerous examples of meta-analysis and research synthesis on mathematics anxiety that were not included, such as the synthesis of 60 years of research (Dowker et al., 2016).

4. Selection of sources

The FOI request showed filtering by peer review and age (since 2010) on some but not all the searches. Gilmore et al. (2021) noted that 82 of the citations came from sources that were not peer reviewed, such as an article from American Educator. However, no articles were included from *Mathematics Teaching*, a non-peer reviewed UK journal of the Association of Teachers of Mathematics, present in Ofsted's data base. Articles from Mathematics Teaching featured frequently in the recreated ERC searches. Far more sources came from the USA than the UK, suggesting selection bias on the part of Ofsted against UK mathematics educators.

Some sources, like Szalontai (2000), were opinion pieces rather than research, although Ofsted stated the reviews would be based on "currently available research evidence" (Ofsted, 2021b, p.3). The statement, "...when lessons and rehearsal opportunities are cut, attainment declines" (Ofsted, 2021a, p.21, footnote 147) was based on this single source that did not include any evidence proving a causal relationship, just Szalontai's opinion.

An analysis of publication dates showed 53% of the references were published within ten years of the review's publication (2021), with a quarter published more than 20 years prior. Older sources can be appropriate when making a historical point, such as the 1939 handbook (footnote 23) or when they are formative works, such as Schoenfeld (1985). However, many of the older sources used in the review were arguably neither important historically nor formative. A statement about systematic approaches

benefitting disadvantaged pupils (footnote 117) was based solely on an overview of Direct Instruction by Kenny (1980), a paper not listed on Kenny's (2023) own website and has only 14 citations in forty years (Google scholar). More recent and significant sources about Direct Instruction appeared in the replicated searches, e.g. a meta-analysis by Stockard et al (2018) but were not included in the review.

5. Citation practices

Citation of texts is a distinguishing feature of research texts, including reviews. The purpose of citation ranges from explicit to implicit. An example of explicit citation practices is the acknowledgement of a source to credit researchers' contributions. Implicit practices include claims to authority or to persuade and are one of the rhetorical tools available to writers (Hyland, 1999).

The section titled 'Quality' included a quotation about textbook use in England (Ofsted, 2021a, p.22, footnote 151) attributed to Brown et al (1998). The statement was made by Bierhoff (1996) but appeared in Brown et al. (1998) because they were challenging it, saying the arguments were "not entirely convincing" (p.368). It was poor research practice for Ofsted to ascribe this statement to Brown et al. rather than Bierhoff; also, they omitted several of the article's authors in their reference list. Askew, one of the acknowledged authors, contacted AMET to contest the use of this statement and went on to indicate that the type of English primary mathematics textbooks in use during the 1990s was likely to be different to those available in 2021, calling into question the inclusion of this statement.

6. Interpretation of research

As Ofsted (2021b) stated, "Educational research is contestable and contested" (p.4). We concur and so interpretation of texts is contestable. However, this does not mean that all

interpretations are equally valid. Epistemologically, the use of citations is part of the social justification through which writers make knowledge claims by positioning their text in wider narratives. In research texts, citation involves invoking the authority of others' research and texts to support these claims.

Many of the sources cited in the review arguably do not support the points made. Comparing the statements in the review with the papers cited, and whether the source provided full, partial or no support for the statement, only 45% of the sources fully supported the points being made, with 26% providing partial support and 28% not supporting. The remaining 1% were links to other documents. In Ofsted's response to AMET, Harford (2021) claimed "it is possible that these concerns may also have arisen as a result of some misinterpretation of the form and function of the mathematics research review" (p.2). This did not address the fact that the table included 19 citations of texts whose authors had challenged Ofsted's interpretation, including [name redacted].

In addition to the issue identified above, Mansell (2022) reported criticism by Askew on the use of Brown et al. (2003) and complaints to Ofsted made by authors such as Gravemeijers and Schoenfeld about the misuse of their work. We give an example of one of these complaints and Ofsted's response. Ofsted's interpretation of the first source (cited in Ofsted, 2021a, p.8, footnote 26) is disputed by Schoenfeld, its author: "All of the references to problem solving in the report, including the specific one that references my foundational research on problem solving, are incorrect". He explained that Ofsted were applying a pre-1985 definition of problem solving that was about exercises for skill development rather than true problems (Mansell, 2022, n.p.). Ofsted's Director's response was that Schoenfeld's work had been chosen because "of numerous references that speak to the nature of domain-specific expertise changing the

way that a problem-solver would ‘see’ and interpret a problem” (Mansell, 2022, n.p.). This, the National Director implied, justifies the citation about the importance of prior knowledge, which demonstrates an approach to selecting literature based on pre-formulated positions rather than fair interpretation.

The section on ‘Methods for more complex measurements and calculations’ (Ofsted, 2021a, p.13, footnotes 80-86) was particularly unsupported by the citations given. This section starts with a claim that formal written methods “minimise the risk of pupils making accidental errors” (Ofsted, 2021a, p.13, footnote 80), which is not supported by either of the two cited papers in this footnote, Anghileri et al (2002) and Woodward (1991). Anghileri et al (2002) challenged this idea, saying that formal algorithms result in many errors due to being applied mechanistically, while Woodward (1991) did not discuss this issue.

There were also factual errors. In talking about cultural factors, the review stated, “This, for example, may be the reason why 75% of Chinese pupils in English schools on free school meals achieved the expected standard in mathematics at key stage 2 in 2019 compared with 44% of their White British counterparts” (Ofsted, 2021a, p.5, footnote 17) but the statistics from this source were 93% versus 60%.

As stated, 26% of citations partially supported claims made. In a section about tasks, Ofsted claimed, “...when physical apparatus is involved, their attention and learning can be compromised” (Ofsted, 2021a, p.22, footnote 164). This was a partial match to Brown et al (2009) because they warn about children potentially becoming distracted by physical manipulatives, but the article describes how to address this, emphasising that, “too much restriction may inhibit or delay their ability to construct the intended transferable, deep understanding of concepts” (Brown et al, 2009, p.162).

In the same section, Ofsted stated,

> Pupils are more likely to engage in disruptive behaviours if they are expected to complete tasks that they have not mastered the component parts of yet. They are more likely to stay on task and be motivated if tasks are achievable (Ofsted, 2021a, p.23, footnote 166).

This footnote related to Gilbertson et al.'s (2008) research which found their participants were on-task more with easy tasks. However, this research was based on only four children, previously identified as having behaviour problems and difficulties with mathematics, whereas the Ofsted statement generalises this to all pupils. Also, there was no discussion of component parts, although the tasks at the lower levels often were component parts for the higher level tasks. The highest level task was labelled 'frustration' and was deliberately designed to be difficult for the pupils, being beyond what they had been taught. The off-task behaviours in the observation protocol included both disruptive ones, such as talking and fidgeting, and non-disruptive ones, such as staring out of the window. The data merely counted off-task behaviours, with no information about which behaviours actually occurred so it cannot be determined if these were disruptive or non-disruptive. Therefore, the source only provided partial support for the statement.

7. Causal claims

The review made unwarranted causal claims numerous times (Gilmore et al, 2021) when small scale observational studies without controls were referenced. Small scale studies by their nature are context specific. For example, Ofsted claim, "...pupils who are more successful develop better learning behaviours" (Ofsted 2021a, p.20, footnote 136), citing Lim (2007). However, Lim's research was an interview and observational study involving five schools to identify characteristics of effective mathematics teaching in Shanghai schools. This type of research design cannot determine causal

relationships.

Lim had not actually made any causal claims. Classroom behaviours and teaching approaches were described only as being "possible contributing factors to high mathematics achievement" (Lim, 2007, p.86). Additionally, Lim (2007) cautioned against adopting approaches from other cultures because cultural differences can mean they will not be as effective in other countries. In general, the relationship between success in mathematics and dispositions such as confidence and learning dispositions is complex and debatable: "Positive attitudes are important, but there is scant evidence on the most effective ways to foster them" (EEF, 2022, n.p.).

8. Generalisation

Ofsted (2021a) provided no discussion about the basis of their claims, how the research was conducted, the type and size of sample, nor the countries involved. Generalisability was implied, even where the authors themselves stated the results should not be generalised.

Discussing low stakes testing Ofsted claimed, "When pupils obtain levels of proficiency, they look forward to and enjoy tests" (Ofsted, 2021a, p.25, footnote 186). This footnote related to a single source, Chiesa and Robertson (2000), about a 12-week programme for a group of five children aged 9 and 10. This sample is far too small to generalise to children of all ages. The claim was presumably based on the statement "...the children in our study took to the procedures enthusiastically and frequently asked for more practice sheets and time probes" (Chiesa and Robertson, 2000, p.308). However, these 'time probes' were not conventional tests; they were tailored to the child and topic but only lasted from a minimum of 5 seconds to a maximum of 5 minutes. This is not the customary meaning of 'test' in England.

The review's sources came from a range of countries but primarily from the United States, East Asia, and Europe. The East Asian sources were sometimes homogenised, implying that practices were the same across the various countries. In the section on 'Equity', one paragraph started by talking about East Asian classrooms and then gave data about pupils in Singapore. This paragraph ended with the statement, "The reason for this success is because a powerful curriculum and plenty of opportunities to engage in purposeful, intelligent practice lead to better outcomes for pupils" (Ofsted, 2021a, p.17, footnote 112). Given the context of the sentence "this success" can be assumed to relate to Singapore specifically, or perhaps to East Asia more generally. However, the two sources used for footnote 112 are Binder and Watkins (1990), an article based in the United States about precision teaching and direct instruction, teaching strategies which are not used in Singapore, and Chen and Li (2010) which is a case study of a single Chinese teacher. Neither of these studies relate to Singapore, and the practice of East Asian countries cannot be generalised from one teacher in China. However, since no details were provided about the sources cited it was not possible for the reader to judge this.

A particular concern about overgeneralisation was identified by The Early Childhood Mathematics Group (ECMG), noting weaknesses with specific reference to Early Years practice. In their response to Ofsted, they identified misapplication of literature and the omission of key literature, particularly related to Early Years. They noted the overgeneralisation to Early Years children of research conducted with older pupils and worried that some of the messages would "undermine the key messages for teachers from research into early mathematics" (ECMG, 2021, p.3).

Problem solving as exemplification

The problematic research approach supported teaching recommendations that depart

from those established by decades of research and practice. One example of this is about teaching for fluency and understanding, generally accepted as both important and interconnected (DfE, 2021; Drury, 2018; EEF, 2022; NCETM 2022; Nunes et al., 2008). The Ofsted review emphasised focussing on fluency; conceptual understanding is not an explicit goal. Rather, Ofsted refer to 'conditional knowledge', meaning knowing when to apply knowledge of procedures and facts. This only encompasses one aspect of conceptual understanding. This contrasts with the English mathematics National Curriculum non-statutory guidance:

> ...the intention should always be to develop students' understanding of mathematical concepts and structures, alongside providing sufficient practice to attain fluency. This combination of developing fluency and mathematical understanding in tandem will enable students to use their learning accurately, efficiently, and flexibly to reason mathematically and solve routine and nonroutine problems, so meeting the aims of the national curriculum (DfE, 2021, p.6).

If the review recommendations were followed it could weaken practice (Watson & Back, 2021). We illustrate these concerns by critiquing Ofsted's view of problem solving. In so doing we expand on concerns raised by Schoenfeld, as well as others (e.g. Tyler, 2021; Watson, A., 2021). Central to this issue is the conflation of problem solving with word problems, ignoring non-routine problems (Tyler, 2021). The Ofsted review frequently refers to word problems and identification of problem types in order to match these to specific strategies but does not mention non-routine problems. Non-routine problems require working out how to tackle the task rather than routine problems where methods are memorised from prior teaching (Burkhardt, 2014, p.3).

Ofsted's restricted view of problem solving, reducing it to word problems with predictable structures, used to apply algorithms differs from established understandings in mathematics focused on novel problems where solution strategies are not immediately obvious (Klerlein & Harvey, 2021): "...not too easy and not mere routine problems; some of them demand originality and ingenuity" (Polya, 1945/2005, p.155). Ofsted's formulaic approach departs from conceptions of problem solving as improvisational, involving risk taking and mistake making (Askew, 2016).

Open-ended problems are mentioned somewhat dismissively: "might be enjoyable for both teachers and pupils (footnote 129) but it does not necessarily lead to improved results (footnote 130) (p.19)." However, Jeffes et al (2013), who are cited in footnote 130 found that students were not being given enough open-ended questions, while the PISA report cited in footnote 128 recommended posing complex, non-routine problems with no obvious solution because these were associated with higher attainment (Burge et al., 2015).

Ofsted's view conflicts with the aims of the National Curriculum for mathematics in England, which talks about "increasingly complex problems" and calls for students to "solve problems by applying their mathematics to a variety of routine and non-routine problems with increasing sophistication" (DfE, 2013, p.3). These aims speak of complexity and unpredictability in problem solving. Ofsted's conception of problems also is in tension with the recommendations from the EEF (2022):

> Select genuine problem-solving tasks that pupils do not have well-rehearsed, ready-made methods to solve. Sometimes problem-solving is taken to mean routine questions set in context, or 'word problems', designed to illustrate the use of a specific method. But if students are only required to carry out a given

procedure or algorithm to arrive at the solution, it is not really problem solving; rather, it is just practising the procedure (p.14).

In advocating for their position on problem solving, Ofsted's misuse of research exemplifies many of the criticisms we made previously. Ofsted (2021a) cites three sources to support the claim that, "Pupils need to be fluent with the relevant facts and methods before being expected to learn how to apply them to problem-solving conditions (footnote 94)" (p.15). A critical word in this statement is "before". A comparison between these sources and the statement is demonstrated in Table 2. The statement has partial support from two of the sources but there is selectivity of evidence and oversimplification. Also, Zhang et al (2018) emphasised that no causal relationships could be asserted from their correlational study.

Table 2. Alignment of sources for Ofsted review footnote 94

Sources listed for footnote 94	**Position**	**Alignment with Ofsted**
National Mathematics Advisory Panel (2008)	Conceptual understanding, fluency, and problem solving need to be developed simultaneously and are mutually reinforcing for pupils from preschool through secondary	Does not support that fluency must come first
Decker and Roberts (2015)	Found basic calculation skills correlated to problem solving but much greater variance accounted for by concept formation, verbal comprehension, and spatial relations Recommended greater emphasis on visual-spatial activities	Partial support: importance of fluency but ignored more significant factors and the recommendations
Zhang et al (2018)	Many factors correlate with solving word problems: fact fluency; working memory; reading comprehension; maths anxiety Did not discuss fluency with methods	Partial support: fact fluency but not methods and ignored other factors and the recommendations

	Recommended analysing children's mistakes to determine which factor was more relevant	

The sources cited support a more nuanced position than Ofsted's. Problem solving tasks are not merely an opportunity to practise the mathematics that has been learned; problem solving involves activities that challenge students and can develop their understanding of mathematics (Askew, 2016; NCTM, 2021).

The suggestion that problem-solving should only come once the facts and methods are already fluent has risks for lower attaining children being excluded from problem solving activities if their memorisation of facts is seen as not secure. This is contrary to evidence of the value for all learners of engaging in challenging open mathematical tasks (e.g., Watson & DeGeest, 2005; Hodgen et al., 2018).

Understanding the review as policy-based evidence

In this section, we draw on our wider policy analysis to consider important issues that can help explain how and why Ofsted's mathematics review was produced and how it is contextualised in current discourses of evidence in education policy. We conclude that the review is policy-based evidence produced to justify or mobilise for a policy position; policy-based evidence contrasts with evidence-based policy (Strassheim & Kettunen, 2014). This may partially explain why the review departed from usual scholarly norms. Whilst evidence-based policy is an open aspiration in policy discourse, policy-based evidence will generally be hidden. Four features of policy-based evidence are:

- Knowledge-monopolisation
- Black-boxing

- Blame-avoidance
- Over-simplification

In the case of the Ofsted review, the rhetoric of evidence-based practice itself serves to obscure how the evidence (the review) follows rather than precedes the policy; Ofsted's (2020) policy prescription for mathematics teaching and learning had already been shared with inspectors.

In Table 3 we summarise Ofsted's policy view contained in the review itself and also supported by Ofsted's own summaries in presentation to teachers (e.g. Ofstednews, 2021). We use categories adapted from a framework used to analyse ideologies of mathematics education (Ernest, 1998). The summary provides an oversight into the positions taken in the review.

Table 3. Ofsted's positions on mathematics and mathematics teaching in the review

Position	**Details**
Theory of mathematics teaching	Teacher-directed explicit instruction Memorisation Retrieval practice Standard formalised methods, some temporary modifications needed in limited cases e.g. younger learners using manipulatives as a step to formal methods
View of mathematics	Procedural, declarative, and conditional knowledge Declarative knowledge as facts precedes conditional knowledge as relationship between facts Algorithmic - a series of steps to follow and "learn by heart"
Theory of learning	Change in long term memory Formal pedagogies apply to all ages (4-18 years old)

	Facts and methods should be memorised before application
View of learners	Novices At risk of cognitive overload and likely to forget, needing memorisation exercises Learners' own methods are unreliable

There are two intertwined reasons that may have led to Ofsted embracing policy-based evidence, even if unintentionally. The first was a change in Ofsted's leadership and direction to more closely align with government ministers' views on education. In 2015, Ofsted's independence from government reduced. Ofsted changed the inspection framework, addressing a previous perceived divergence from government educational ideology, with the Schools Minister referring to the previous inspection approach as "Ofsted's reign of error" (Gibb, 2015). During this 'reign of error', Ofsted's research into teaching and learning in English schools focused on more inquisitorial and meditative (reflective) approaches with reports on practices in the various curriculum subjects. This was represented in mathematics in a number of reports (e.g. Ofsted, 2012) that sought to reflect back to schools examples of practices that had broad consensus amongst teachers, teacher subject associations, researchers and teacher educators as being fruitful for mathematics learning (Compton & Boylan, 2023). A key feature was to embrace subject specificity rather than, as recently, a more generic view of learning and teaching.

Secondly, Ofsted embraced the discourse of evidence, with the Ofsted chief inspector stating "we aim to ground all our work in research evidence" (Spielman, 2023). The revised inspection framework (Ofsted, 2019a) was informed by a review of research (Ofsted, 2019b) which in turn informed principles that underpinned the research reviews (Ofsted, 2021b), into subjects including mathematics. Ofsted formed a

‘Curriculum Unit’ that had an important role in the development of the Inspection Framework, the subject review methodology, and the subject reviews themselves. Members of both the Unit and subject advisory groups appear to be recruited from educators sympathetic to government education ideology and a ‘traditionalist’ view of education (Watson, S., 2021).

These changes represented an ideological turn by Ofsted in relation to teaching and curriculum, as well as to the use of evidence. Ofsted (2021b) stated that the research reviews would use research that aligned with their established conceptions of subject quality. This indicates that they were providing policy-based evidence, unlike, for example, the NRC which was seeking to develop evidence-based policy.

The revised inspection framework placed less emphasis on school performance data, such as examination outcomes, but rather focused on schools’ curriculum practices (Ofsted, 2019a), emphasising the importance of a knowledge rich curriculum and teaching informed by supposed insights from cognitive science. This aligned with other government education policies such as the Core Content Framework for Initial Teacher Education (DfE, 2019a) and the Early Career Framework for newly qualified teachers (DfE, 2019b). The internal inspection guidance documents for primary and secondary mathematics (Ofsted, 2020), published the year before the review and which tell inspectors what to look for in their ‘deep dives’ into mathematics teaching in schools, demonstrate their positionality. For example,

> ...demonstrating proof of ‘understanding’ will not guarantee that pupils learn useful facts, methods and strategies. Moments of understanding, no matter how powerful, are likely to be fleeting. (Ofsted, 2020, p.12)

This emphasis on facts over understanding was evident in several sections of the review, although stated less starkly and contrasts with the stated aims of the National Curriculum (DfE, 2013).

In departing from usual epistemic practices of research, the mathematics review does not translate or broker knowledge. Rather it produces it, and so expands Ofsted's epistemic role in mathematics education, and represents a departure from Ofsted's historical relationship to research and evidence (Smith, 2000). Ofsted claimed greater authority to decide what researchers' texts mean than the researchers themselves (as indicated by responses to researchers' complaints). It exemplifies policy-based evidence, that is the production of evidence to justify or mobilise for a policy position. Specifically, it uses mechanisms previously identified (Strassheim & Kettunen, 2014).

in the production of policy-based evidence:

- oversimplification of claims
- knowledge monopolisation in not basing recommendations in systematic reviews such as EEF or alternative rigorous expert reviews
- black boxing in the lack of transparency necessitating a Freedom of Information Request

Given this, Ofsted have moved beyond a regulative role, inspecting how policy is enacted, to a policy-making role. This was exemplified earlier in the discrepancies between the review's approach to problem solving and the National Curriculum's aims and guidance (DfE, 2013).

Conclusion

We have argued that across eight features, the Ofsted mathematics review was not undertaken with sufficient attention to review quality. Ofsted's role, influence and recent ideological turn suggests the mathematics review, and other subject reviews, are

performative documents - essentially rhetorical texts (Craske, 2021) rather than scholarly outputs. We have acknowledged that Ofsted, in response to criticism, have suggested that the review is better seen as a position paper and we would agree.

However, the use of the rhetoric of evidence and research obscures the ideological nature of at least some of Ofsted's recommendations, while the lack of criticality in the discussion of the evidence promotes the idea that evidence cannot be questioned. In appropriating and misusing the discursive practices of research and review, Ofsted use their warrant to claim certain approaches to mathematics teaching are compelled by evidence. In this they are using policy-based evidence.

We have argued the review was undertaken to support a pre-existing position. We acknowledge that the same critique could be made of our critical response. Our positionality, as academics, and our different view of quality in mathematics education influences our analysis and conclusions, just as Ofsted's review was influenced by theirs. The interpretation of sources is particularly prone to this subjectivity, which is why we sought confirmation from the original authors, where possible. We have also included quotations so that readers can make their own judgements.

In our view, Ofsted researchers were set an unreasonable task, to undertake a technically, methodologically demanding wide-ranging academic review without sufficient resources to do this. More positively, the Ofsted review did embrace sources using a wide range of research methods, including small-scale qualitative studies. Notwithstanding our concerns about causal claims and over generalisation, the use of a broader range of research than is sometimes included in systematic reviews is welcome.

An alternative approach to Ofsted undertaking its own reviews would be to engage with, translate and broker evidence from existing reviews of various types (e.g., Nunes et al., 2008; Hodgen et al., 2018). We question whether undertaking reviews of

research is appropriate or efficient for an inspectorate, given the capacity elsewhere in the system in England and internationally through various specialist bodies.

Since the publication of the review, under new leadership and a new government (Roberts and Maisuria, 2024) Ofsted has undergone significant reforms. As part of these, in relation to research activities, we believe Ofsted should reconsider its role as part of a wider evidence system. In our discussion of types of knowledge inspectorates produce, we drew on a three-fold framework of regulative, inquisitorial, and meditative knowledge (Jacobsson, 2006) and extended this to consider Ofsted's role as a knowledge broker. In our view, Ofsted has an important role in using its inquisitorial knowledge of current practice in the system to produce reports (e.g. Ofsted, 2023); indeed, inspection services are uniquely placed for this role.

Although our paper is substantially a critique of Ofsted's review, we note the research community is also implicated in deficiencies in the intersection of research and evidence, policy and education practice. Research producers and evidence brokers need to consider how to ensure that research informed guidance is (even more) accessible to inspectors and teachers to enable them to develop and enact research-based policy. Further, our analysis of Ofsted's review was challenged by the lack of agreed criteria for quality for reviews when a systematic review is not appropriate or feasible. Consequently, we developed a framework of features. Although additional to our main purpose in this paper, we propose that these features potentially offer an analytical framework for analysis of quality for other reviews in mathematics education and beyond. This framework could help support teachers and school leaders to assess claims, made by official bodies and others, based upon reviews of evidence.

Of the eight features we identified in relationship to review quality, we conclude by emphasising one of these as important more generally to evidence discourses and

practices in education internationally – transparency. Transparency also serves to counter the risk of 'black boxing' implicated in policy-based evidence. Given the political and contested nature of both education and research evidence, transparency is not only important about methods but also about the meanings ascribed to 'evidence' and 'evidence-based', acknowledging that these terms are contested and political. Given, mathematics education internationally is influenced by discourses of evidence-based practice, it is important to be able to analyse how these have been used to influence educational policy and be aware of the politicisation of such evidence. Acknowledging that there are different sources of evidence and evidence has multiple meanings can also serve to mitigate the risk of knowledge monopolisation if only a restricted set of practices in mathematics education are seen as 'evidence-based'. Similarly, understanding the meaning of evidence as contested and multiple can support greater understanding of the affordances and limitations of the different types of research and evidence that can inform mathematics teaching and mathematics education policy.

Disclosure statement: there are no competing interests to declare.

Acknowledgements: We are grateful to the journal editors for the suggestion that an article on the Ofsted review would make a useful contribution to literature on mathematics education and policy. We benefited greatly from insightful and constructive reviews by both anonymous reviewers and editorial comments.

References

Alexander, P. A. (2020). Methodological guidance paper: The art and science of quality systematic reviews. *Review of Educational Research, 90*(1), 6-23. https://doi.org/10.3102/0034654319854352

Anghileri, J., Beishuizen, M. & Van Putten, K. (2002). From informal strategies to structured procedures: Mind the gap! *Educational Studies in Mathematics, 49*, 149-170. https://doi.org/10.1023/A:1016273328213

Askew, M. (2016). *Transforming primary mathematics*. Routledge.

Askew, M., & Wiliam, D. (1995). *Recent research in mathematics education 5 – 16.* HMSO.

Baxter, J., & Hult, A. (2017). Different systems, different identities: The work of inspectors in Sweden and England. In J. Baxter (Ed.), *School inspectors: Policy implementers, policy shapers in national policy contexts*. (pp.45-69). Springer.

Benn, M. (2023). Inside the 'new educational establishment' in England. In V. Ellis (Ed.) *Teacher education in crisis: The state, the market and the universities in England.* (pp. 27-31). Bloomsbury.

Bierhoff, H. (1996). Laying the foundations of numeracy: A comparison of primary school textbooks in Britain, Germany and Switzerland. *Teaching Mathematics and Its Applications, 15(*4), 141-160. https://doi.org/10.1093/teamat/15.4.141

Binder, C., & Watkins, C. (1990). Precision teaching and direct instruction: measurably superior instructional technology in schools, *Performance Improvement Quarterly, 3*(4), 74–96. https://doi.org/10.1111/j.1937-8327.1990.tb00478.x

Boylan, M., & Adams, G. (2023). Market mirages and the state's role in professional learning: the case of English mathematics education. *Journal of Education Policy*, *39*(2), 1-23. https://doi.org/10.1080/02680939.2023.2195854

Brown, M., Askew, M., Baker, D., Denvir, H., & Millett, A. (1998). Is the national numeracy strategy research-based? *British Journal of Educational Studies*, *46*(4), 362–385. https://doi.org/10.1111/1467-8527.00090

Brown, M., Askew, M., & Millett, A. (2003). How has the national numeracy strategy affected attainment and teaching in year 4? In J. Williams (Ed.), *Proceedings of the British society for research into learning mathematics*, *23*(2), 13-18.

Brown, M., McNeil, N., & Glenberg, A. (2009). Using concreteness in education: real problems, potential solutions. *Child Development Perspectives*, *3*(3), 160-164. https://doi.org/10.1111/j.1750-8606.2009.00098.x

Burge, B., Lenkeit, J., & Sizmur, J. (2015). *PISA in practice: Cognitive acceleration in maths*. NFER. https://www.nfer.ac.uk/media/vmtj0jne/pquk04.pdf

Burkhardt, H. (2014). *Curriculum Design and Systemic Change*. Shell Centre. https://www.mathshell.com/papers/pdf/hb_2013_yp_gl_curric_chapter.pdf

Carey, E., Devine, A., Hill, F., Dowker, A., McLellan R., & Szucs, D. (2019). *Understanding maths anxiety.* Nuffield Foundation. https://www.nuffieldfoundation.org/project/understanding-mathematics-anxiety

Chen, X., & Li, Y. (2010). Instructional coherence in Chinese mathematics classroom—a case study of lessons on fraction division. *International Journal of Science and Mathematics Education*, *8*(4), 711–735. https://doi.org/10.1007/s10763-009-9182-y

Chiesa, M., & Robertson, A. (2000). Precision teaching and fluency training: making maths easier for pupils and teachers. *Educational Psychology in Practice*, *16*(3), 297–310. https://doi.org/10.1080/713666088

Coldwell, M. (2022). Evidence-informed teaching in England. In C. Brown and J. Malin (Eds.) *The Emerald Handbook of Evidence-Informed Practice in Education.* (pp. 59-68). Emerald Publishing Limited.

Compton, A. (2021). Reviewing the review. *Mathematics Teaching, 278*, 23. https://www.atm.org.uk/write/MediaUploads/Journals/MT278/06.pdf

Compton, A. & Boylan, M. (2023). From flexible to restricted mathematics: 15 years of Ofsted mathematics reports. *Mathematics Teaching, 289*, 31-36.

Craske, J. (2021). Logics, rhetoric and 'the blob': Populist logic in the Conservative reforms to English schooling. *British Educational Research Journal*, *47*(2), 279-298. https://doi.org/10.1002/berj.3682

Decker, S. & Roberts, A. (2015). Specific cognitive predictors of early math problem solving, *Psychology in Schools*, *52*(5), 477-488. https://doi.org/10.1002/pits.21837

DfE. (2013). National curriculum in England: mathematics programmes of study. https://www.gov.uk/government/publications/national-curriculum-in-england-mathematics-programmes-of-study

DfE. (2019a). ITT Core content framework. https://assets.publishing.service.gov.uk/media/6061eb9cd3bf7f5cde260984/ITT_core_content_framework_.pdf

DfE. (2019b). Early career framework. https://www.gov.uk/government/publications/early-career-framework

DfE. (2021). Mathematics guidance: Key Stage 3 Non-statutory guidance for the national curriculum in England https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1056795/KS3_NonStatutory_Guidance_Sept_2021_FINAL_NCETM.pdf

Dowker, A., Sarkar, A., & Looi, C.Y. (2016). Mathematics anxiety: What have we learned in 60 years? *Frontiers in Psychology*, *7*(508) doi: 10.3389/fpsyg.2016.00508 https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00508/full

Drury, H. (2018). *How to teach mathematics for mastery.* Oxford University Press-Children.

Early Childhood Maths Group. (2021). *Response to the Ofsted mathematics review of research.* https://earlymaths.org/response-to-the-ofsted-mathematics-review-of-research/

EEF. (2022). *Improving Mathematics in Key Stages 2 and 3: Eight recommendations to improve outcomes in maths for 7–14-year-olds.* https://educationendowmentfoundation.org.uk/education-evidence/guidance-reports/maths-ks-2-3

Ernest, P. (1998). *The philosophy of mathematics education*. SUNY press.

Fearn, H. [@HeatherBellaF]. (2021, July 12). *We did not want people to think they were literature reviews in the true sense hence not calling them that.* [Tweet]. Twitter. https://twitter.com/HeatherBellaF/status/1414676448632311809

Gibb, N. (2015, September 5). *Nick Gibb: The importance of the teaching profession*, (Speech). https://www.gov.uk/government/speeches/nick-gibb-the-importance-of-the-teaching-profession

Gilbertson, D., Duhon, G., Witt, J. C., & Dufrene, B. (2008). Effects of academic response rates on time-on-task in the classroom for students at academic and behavioral risk. *Education & Treatment of Children*, *31*(2), 153–165. https://dx.doi.org/10.1353/etc.0.0020.

Gilmore, C., Trundley, R., Bahnmueller, J., & Xenidou-Dervou, I. (2021). How research findings can be used to inform educational practice and what can go wrong: The Ofsted Mathematics Research Review 2021. *Mathematics Teaching*, *278*, 35-38. https://www.atm.org.uk/write/MediaUploads/Journals/MT278/12.pdf

Gustafsson, J. E., Ehren, M. C., Conyngham, G., McNamara, G., Altrichter, H., & O'Hara, J. (2015). From inspection to quality: Ways in which school inspection influences change in schools. *Studies in educational evaluation*, *47*, 47-57. https://doi.org/10.1016/j.stueduc.2015.07.002

Harford, S. (2021, 6 August). Correspondence regarding mathematics research review. https://www.ametonline.org.uk/app/download/12867529/Ofsted+response+to+AMET+060821.pdf

Helgetun, J., & Menter, I. (2022). From an age of measurement to an evidence era? Policy-making in teacher education in England. *Journal of Education Policy*, *37*(1), 88-105. https://doi.org/10.1080/02680939.2020.1748722

Hodgen, J., Foster, C., Marks, R., & Brown, M. (2018). *Improving mathematics in key stages two and three: Evidence review.* Education Endowment Foundation (EEF). https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/mathematics-in-key-stages-2-and-3

HCST (House of Commons Select Committee on Technology). (2018). *Research integrity: Government and UK Research and Innovation Responses to the Committee's Sixth Report*. https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/1563/1563.pdf

Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied linguistics*, *20*(3), 341-367. https://doi.org/10.1093/applin/20.3.341

Jacobsson, B. (2006). Regulated regulators. Global trends of State Transformation. In M-L. Djelic & K. Sahlin-Andersson (Eds.), *Transnational Governance: Institutional Dynamics of Regulation* (pp. 205-224). Cambridge.

Jeffes, J., Jones, E., Wilson, M., Lamont, E., Straw, S., Wheater, R., & Dawson, A. (2013). *Research into the impact of Project Maths on student achievement, learning and motivation*. NFER. https://www.nfer.ac.uk/media/215dng42/research_into_the_impact_of_project_maths_on_student_achievement_learning_and_motivation.pdf

Kenny, D. (1980). Direct instruction: an overview of theory and practice, *Journal of the Association of Special Education Teachers*, *15*, 1-6.

Kenny, D. (2023). *Professor Dianna T. Kenn: Articles & videos.* https://diannakenny.com.au/blog/

Klerlein, J., & Hervey, S. (2021). *Mathematics as a Complex Problem-Solving activity*. Generation Ready. https://www.generationready.com/white-papers/mathematics-as-a-complex-problem-solving-activity/

Lim, C. (2007). Characteristics of mathematics teaching in Shanghai, China: through the lens of a Malaysian, *Mathematics Education Research Journal*, *19*(1), 77-88. https://doi.org/10.1007/BF03217450

Mansell, W. (2022). *Ofsted's maths "research review" prompted complaints from renowned international academics that their findings had been misinterpreted*. Education Uncovered. https://www.educationuncovered.co.uk/news/162056/ofsteds-maths-research-review-prompted-complaints-from-renowned-international-academics-that-their-findings-had-been-misinterpreted.thtml

McVeigh, H. (2020). *Teaching and the role of Ofsted*. UCL IoE press.

Murphy, P., Dowd, A., Lloyd, G., & List, A. (2020). Transparency in Literature Syntheses and Editorial Review: Introducing the Methodological Guidance Paper Series. *Review of Educational Research,* 90(1), 3-5. https://doi.org/10.3102/0034654319901128

National Mathematics Advisory Panel. (2008). *Foundations for success: the final report of the National Mathematics Advisory Panel*. US Department for Education.

National Research Council. (2001). Adding it up: Helping children learn mathematics. J. Kilpatrick, J. Swafford, and B. Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. National Academy Press.

NCTM. (2021). *Problem solving*. https://www.nctm.org/Research-and-Advocacy/Research-Brief-and-Clips/Problem-Solving/

NCETM. (2022). *The essence of teaching mathematics for mastery.* https://www.ncetm.org.uk/teaching-for-mastery/mastery-explained/the-essence-of-mathematics-teaching-for-mastery/

Nunes, T., Bryant, P., & Watson, A. (2008). *Paper 8: Methodological appendix*. Nuffield Foundation. https://www.nuffieldfoundation.org/wp-content/uploads/2019/11/P8.pdf

Ofsted. (2012) Made to measure: Messages from inspection evidence. https://assets.publishing.service.gov.uk/media/5a8163dde5274a2e87dbd5f7/Mathematics_made_to_m easure.pdf

Ofsted. (2019a). *Education Inspection Framework (EIF).* https://www.gov.uk/government/publications/education-inspection-framework

Ofsted. (2019b). Education inspection framework: Review of evidence. https://assets.publishing.service.gov.uk/media/6034be17d3bf7f265dbbe2ef/Research_for_EIF_framework_updated_references_22_Feb_2021.pdf

Ofsted. (2020). *Inspector guidance: Primary mathematics.* Quality Schools. https://www.quality-schools.com/ofsted-inspector-subject-crib-sheets-what-ofsted-didnt-want-you-to-know/

Ofsted. (2021a). *Research review series: mathematics. A review of research into factors that influence the quality of mathematics education in schools in England*. https://www.gov.uk/government/publications/research-review-series-mathematics

Ofsted. (2021b). *Principles behind Ofsted's research reviews and subject reports.* https://www.gov.uk/government/publications/principles-behind-ofsteds-research-reviews-and-subject-reports/principles-behind-ofsteds-research-reviews-and-subject-reports

Ofsted. (2023). *Coordinating mathematics success. Mathematics subject report.* www.gov.uk/government/publications/subject-report-series-maths/coordinating-mathematical-success-the-mathematics-subject-report

Ofstednews. (2021 November 3). *Mathematics: Webinar from North West region*. [Video]. YouTube. https://www.youtube.com/watch?v=QK9vAAIWiCM&list=PLLq-zBnUkspPdslDqP1pwcYY0DBlGVbgl&index=6

Ozga, J. (2014). Knowledge, inspection and the work of governing. *Sisyphus—Journal of Education*, 2(1), 16-38.

- Perryman, J., Maguire, M., Braun, A., & Ball, S. (2018). Surveillance, governmentality and moving the goalposts: The influence of Ofsted on the work of schools in a post-panoptic era. *British Journal of Educational Studies*, *66*(2), 145-163. https://doi.org/10.1080/00071005.2017.1372560

Pólya, G. (2004). *How to solve it: a new aspect of mathematical method*. Princeton University Press. (Original work published 1945)

PRISMA. (2024). Transparent reporting of systematic reviews and meta-analysis. http://www.prisma-statement.org/

Reynolds, D. (1996). *Worlds apart? A review of international surveys of educational achievement involving England*. Ofsted.

Rizvi, F. & Lingard, B. (2010). *Globalizing education policy*. Routledge.

Roberts, N., & Maisuria, A. (2024). Ofsted: School inspections in England. The Commons Library. https://researchbriefings.files.parliament.uk/documents/SN07091/SN07091.pdf

Science of Math. (n.d.). Science of Math. https://www.thescienceofmath.com/

Schoenfeld, A. (1985). *Mathematical problem solving*. Academic Press.

- Smith, G. (2000). Research and inspection: HMI and OFSTED, 1981-1996-a commentary. *Oxford Review of Education*, *26*(3-4), 333-352. https://doi.org/10.1080/713688535

Strassheim, H., & Kettunen, P. (2014). When does evidence-based policy turn into policy-based evidence? Configurations, contexts and mechanisms. *Evidence & Policy*, *10*(2), 259-277. https://doi.org/10.1332/174426514X13990433991320

Spielman A. (2023). *Amanda Spielman's speech to Oxford University's Department of Education,* (Speech). https://www.gov.uk/government/speeches/amanda-spielmans-speech-to-the-university-of-oxfords-department-of-education.

Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, *88*(4), 479-507. https://doi.org/10.3102/0034654317751919

Suri, H., & Clarke, D. (2009). Advancements in research synthesis methods: From a methodologically inclusive perspective. *Review of Educational Research*, *79*(1), 395-430. https://doi.org/10.3102/0034654308326349

Sutton, A., Clowes, M., Preston, L. & Booth, A. (2019). Meeting the review family: exploring review types and associated information retrieval requirements. *Health Information and Libraries Journal* 36 (3), 202–222. https://doi.org/10.1111/hir.12276

Szalontai, T. (2000). *Some facts and tendencies in Hungarian mathematics teaching*. Institute of Mathematics and Informatics. https://www.cimt.org.uk/journal/tshungmt.pdf

Tyler, L. (2021). *Ofsted's research review: standing on the shoulders of (maths) giants.* Ark Curriculum Plus. https://www.arkcurriculumplus.org.uk/news-events/ofsteds-research-review-standing-on-the-shoulders-of-maths-giants

Watson, A. (2021). *Anne's response to Ofsted "research review: mathematics"*. Promoting Mathematical Thinking. http://www.pmtheta.com/

Watson, A., & Back, J. (2021, October 18). Ofsted's maths research review could lead to poorer teaching. *Schools Week*. https://schoolsweek.co.uk/ofsteds-maths-research-review-could-lead-to-poorer-teaching/

Watson, A., & de Geest, E. (2005). Principled teaching for deep progress: Improving mathematical learning beyond methods and materials. *Educational studies in mathematics, 58*, 209-234. https://doi.org/10.1007/s10649-005-2756-x

Watson, S. (2021). New Right 2.0: Teacher populism on social media in England. *British Educational Research Journal*, *47*(2), 299-315. https://doi.org/10.1002/berj.3664

West, R., Stenius, K., & Kettunen, T. (2017). Use and abuse of citations. In T. Babor, K. Stenius, R. Pates, M. Miovský, J. O'Reilly & P. Candon. *Addiction Science: A Guide for the Perplexed*. (pp. 191-206). Ubiquity Press.

Wadhwa, M., Zheng, J., & Cook, T. D. (2024). How consistent are meanings of "evidence-based"? A comparative review of 12 clearinghouses that rate the effectiveness of educational programs. *Review of Educational Research*, *94*(1), 3-32. https://doi.org/10.3102/00346543231152262

Woodward, J. (1991). Procedural knowledge in mathematics: The role of the curriculum. *Journal of Learning Disabilities. 24*(4), 242-251. https://doi.org/10.1177/002221949102400408

Zhang, J., Cheung, S., Wu, C. & Meng, Y. (2018). Cognitive and affective correlates of Chinese children's mathematical problem solving, *Frontiers in Psychology*, *9* (2357), 1-10. doi: 10.3389.