What makes a good educational research summary? A comparative judgement study of mathematics teachers' and mathematics education researchers' views

Lucy Rycroft-Smith | Andreas J. Stylianides

Faculty of Education, University of Cambridge, Cambridge, UK

Correspondence
Lucy Rycroft-Smith, University of Cambridge, Faculty of Education, 184 Hills Rd, Cambridge CB2 8PQ, UK.
Email: lr444@cam.ac.uk

Funding information
None.

Abstract
The field of knowledge brokering in education—aiming to better connect research to practice—is currently emerging. Evidence of a community dissonance between researchers and practitioners in education suggests that models of knowledge brokering that consider the perspectives and priorities of both groups are required. It is also a priority to identify what kinds of knowledge brokering products such as research summaries are successful at communicating research to teachers, potentially functioning as boundary objects. We report findings from a comparative judgement study where a group of 28 mathematics teachers and a group of 19 mathematics education researchers ranked twenty research summaries, from different sources, in terms of their success at communicating the research to teachers, and explained which features contributed to this ranking. Overall the findings suggest moderate consensus both within and between the two groups, and some important areas of difference. Both groups agreed that graphic design was the most important
element of a research summary; that being summarised, having implications for practice and being easy to read or accessible and well-structured were key features of a research summary; and that the length of the research summary or the time it might take to read were also important. Whereas teachers mentioned the choice of topic of the research summary and to some extent language more than researchers, researchers highlighted some other features of successful research summaries for them that teachers did not: ideas around methodology and use of evidence, opportunities for critical reflection, and issues of trust and credibility.

**KEYWORDS**
comparative judgement, evidence-into-use, knowledge brokering, mathematics education

**Context and implications**

**Rationale for this study**
The field of knowledge brokering is new and poorly understood. This study provides much-needed empirical evidence on teachers’ and researchers’ views of research summaries.

**Why the new findings matter**
The findings of this study contribute to the literature around knowledge brokering to address community dissonance in educational research and practice, suggesting both areas of agreement and difference between teachers and researchers, and prompting further questions around the use of research summaries as boundary objects.

**Implications for researchers, policy makers and practitioners**
There are important implications for the design of educational research summaries, as both teachers and researchers agreed that graphic design was the most important element of a research summary, and that being summarised, having implications for practice and being easy to read or accessible and well-structured were also all key. The areas where there was less overlap are recommended for further investigation. There are also implications for the use of comparative judgement in new ways, for example in investigating knowledge brokering and educational professional development.
WHAT MAKES A GOOD EDUCATIONAL RESEARCH SUMMARY?

INTRODUCTION

One of the most persistent problems in education is the movement of research-into-use, sometimes characterised as the evidence-to-practice pipeline, the need for implementation of research evidence into practice, or knowledge mobilisation. Researchers tend to agree that the use of educational research knowledge in practice is beneficial to decision-making at all levels but that this does not happen enough (e.g. Farley-Ripple & Grajeda, 2020; Godfrey & Brown, 2019; Stenhouse et al., 1985). Research knowledge is produced in particular education spaces, but it does not appear to move to other spaces—those inhabited by the ‘user’—very often or very consistently (Moss, 2017).

While these issues have been written about for several decades (e.g. Dagenais et al., 2012; Oancea, 2005), and many reasons considered as to why this may be the case, very little is actually known about effective models of such movement—either into practice or policy, which are often written about together:

Helping to embed secure research findings into policy and practice has been a concern for at least as long as the concern about improved evidence itself and is worldwide... We need to know more about how primary research evidence is best channelled to users, so that they are more likely to use evidence-based approaches, and do so because they are evidence-based. (Gorard, See & Siddiqui, 2020, pp. 571-572).

The blockages to evidence use in practice for teachers described in the literature are: language (Ion & Lucu, 2014; van Schaik et al., 2018; See et al., 2016); timeliness (Hering, 2016); finding/accessing research (Farley-Ripple & Grajeda, 2020; Levin, 2013; Martinovic et al., 2012); understanding/interpreting research (Farley-Ripple & Grajeda, 2020); cost (Drill et al., 2013); technology (Drill et al., 2013); having time/energy to devote to research (Drill et al., 2013); lack of personal contact with researchers (Hering, 2016); finding research useful or relevant (Farley-Ripple & Grajeda, 2020; Hering, 2016); being able to apply research ideas in context (Whitehurst, 2003); institutional culture (Hemsley-Brown & Sharp, 2003); and having access to good-quality syntheses or reviews of research (Cordingley, 2008; Farley-Ripple & Grajeda, 2020). Some have suggested that researchers can help solve these problems and need to engage teachers in their research better (e.g. Rickinson, 2016); others have suggested that some responsibility also lies with teachers, who need to spend more time getting to grips with research and building their research literacy skills (e.g. Williams & Coles, 2007).

More recently it has been suggested that the activity of knowledge brokering by experts may help to bridge the gap between research and practice, not just by moving research but also by a range of other processes and activities that may take into account the complexities and intersections of the barriers involved, for example examining research use as a system property involving production, transformation and implementation (Lysenko et al., 2016).

This activity may include creating resources to summarise ideas from research, convening partnerships and dialogue between researchers and teachers, and/or supporting teachers to engage with research (Bush, 2017 in Farley-Ripple & Grajeda, 2020). Because we (as a field) know some of the barriers to users reading and using research, these ideas should contribute towards the effective design of knowledge brokering resources such as research summaries (Grimshaw et al., 2012). However, there is little research so far about what constitutes a good research summary and how researchers’ and practitioners’ perceptions of a successful research summary may relate to one another.

This research study took a step towards addressing these issues, considering in particular the views of mathematics teachers and mathematics education researchers—whether
they agreed within and across groups as to what a successful research summary for teachers may look like. Specifically, we aimed to address the following research questions:

1. Is there consensus within, and across, mathematics teachers and mathematics education researchers on which research summaries they judge to be more and less successful at communicating the research to teachers?
2. Which features emerge as important to mathematics teachers, and to mathematics education researchers, when they consider which research summaries are more and less successful at communicating the research to teachers?

BACKGROUND

Terminology and models of research-into-use

This is a new field, and also a transdisciplinary one, which means that terms are not yet consistently defined or used. However, the profusion of different terminologies and the concomitant metaphors and imageries they conjure give us a useful window as to the ways in which different researchers may perceive the processes of research movement into the spaces of practice.

For example, the word diffusion has been used to refer to the way that research knowledge moves towards and within the teaching profession (Williams & Coles, 2007, p. 185), suggesting multi-directionality and a kind of osmosis-type flow model. However, this may also suggest a particular ideal ratio of research knowledge and ‘other’ knowledges which may emerge after a period of settling down, as well as a sense of randomness or lack of intentionality which belies the expertise and work inherent in the movement of research into practice in useful ways. Other researchers use the phrase knowledge translation; for example: ‘research alone does not inform practice, rather a process of knowledge translation is required to enable research findings to become meaningful for practitioners in their contextual settings’ (Jones, Procter et al., 2015, p. 255). This phrase suggests a process of movement from one context to another, with an intention to preserve meanings intact, and with the role of interpretation, selection of information and restructuring minimised.

Others have used terms which highlight these issues more fully; for example Moss (2017, p. 237) described ‘the troubled routes that knowledge takes into practice’ in education as analogous to the field of medicine, using knowledge mobilisation as a cross-disciplinary term, defined as ‘the complex social processes that underpin the cooption, uptake and use within one sphere of practice of specialised knowledge created in another’ (p. 237). Moss clarifies the use of ‘knowledge mobilisation’ specifically over ‘knowledge transfer’:

[K]nowledge mobilisation is not just about moving a clearly defined set of ideas, concepts, research techniques or information from here to there. Rather it is about grappling with which forms of knowledge are apt in which contexts and how they can be strengthened through use. Where the term ‘knowledge transfer’ obscures these issues, the term ‘knowledge mobilisation’ brings them into focus. (Moss, 2017, p. 238)

This is an idea which appears in other researchers’ work. For example, it has been suggested that there is a risk that research ideas may be communicated to teachers in a linear manner, without space for critical engagement, dialogue or co-creation of knowledge (H. Davies et al., 2008). Considering this, it is clear that using terms such as ‘translation’ or ‘transfer’ may suggest a kind of lack of agency on behalf of the teachers to critique, question and even reject
What makes a good educational research summary?

Research evidence, which is potentially in conflict with their professional identity and may feed into a view of research evidence which erases or devalues other types of knowledge (Rycroft-Smith & Macey, 2021).

Similarly, terminology around teacher use of research reflects ideas about what research could, can and should be used for. A distinction is drawn in many researchers’ work between instrumental, conceptual and symbolic use of research:

- Instrumental use: transmissive applications of research to concrete practice
- Conceptual/enlightenment use: a change in understanding that affects practice more indirectly
- Symbolic/strategic use: justification, influence or support of actions/decisions (Lysenko et al., 2016)

Estabrooks (1999) suggests ‘symbolic’ should be named ‘persuasive’, which alludes to the need to use research in the service of argument in educational contexts, lending legitimacy or the weight of power to a particular position; yet this research is in the context of medicine. We still know very little about the ends to which practitioners apply research in their teaching practice (Dagenais et al., 2012) and therefore what they may need and want from knowledge brokering.

Thus we use the term knowledge brokering in this study to reflect the type of complex, multidirectional and specialised work around knowledge flow that is defined to explicitly consider fostering relationships, promoting mutual understanding, building capacity and boundary spanning (e.g. Conklin et al., 2013) to which we add taking account of criticality.

Community dissonance and knowledge brokering

Since the issue of the research-practice gap appears to persist despite some agreement that evidence use in practice (by teachers and other school-based professionals) would be beneficial, the idea of community dissonance between researchers and teachers has been used (Farley-Ripple et al., 2018). According to Farley-Ripple et al. (2018, p. 237), community dissonance (a term they attribute to Bogenschneider et al., 2010) describes the problem whereby ‘the cultures, contexts, and systems in which researchers and practitioners operate, including institutional goals and professional norms and expectations, differ significantly’. This chasm between research and practice communities in education has also been described as having at its root ‘deep and thorny social, cultural and structural divides’ (Farley-Ripple & Grajeda, 2020, p. 2). Since there is an argument that (at least some) educational research should be relevant and usable—including being focused on matters teachers can actually control (Stoll, 2012)—an understanding of the needs of teachers by researchers is implied as necessary. One barrier to this has been suggested: hierarchical ideas of academic superiority may impair research-practice relationships (Farley-Ripple & Grajeda, 2020; Lysenko et al., 2014), and language is one way in which this has been expressed; the outputs of academic research are often written in technical language that is not accessible or useful to non-academic audiences (Phipps et al., 2020; See et al., 2016) and therefore preserve or gatekeep the separation of the two communities.

However, it is important to note that teachers can be researchers too, and research is not something that is always done separately from or to teachers. For example, there is a large body of work around design-based research (DBR) in education, which involves a collaborative partnership between researchers and teachers (Anderson & Shattuck, 2012). DBR ‘seeks to increase the impact, transfer, and translation of education research into improved practice’ and ‘stresses the need for theory building and the development of design
principles that guide, inform, and improve both practice and research in educational contexts’ (Anderson & Shattuck, 2012, p. 16). More generally, there exists much literature on the blurred boundary between the teacher and researcher roles and the power structures that are involved, with some calls for teachers to engage in the doing of research themselves as part of a process of reprofessionalisation, to ‘challenge the reductionist and technicist methods that promote a “top down” system of education’ (Kincheloe, 2012, p. 3).

There is evidence of an international effort to strengthen connections between educational research and practice through breaking down the community dissonance between teachers and researchers (e.g. Farley-Ripple & Grajeda, 2020), but this is not an easy problem to solve, because it involves changes in various aspects of what has been termed the ‘education ecosystem’, a complex network of systems of actors and agents characterised by external pressures and ongoing change (e.g. Godfrey & Brown, 2019; Hudson, 2017). Hutchinson and Huberman (1994) considered education in itself a ‘universe’, conveying size, complexity and significant boundaries with spaces (and times). Examination of some of these boundaries and connections of this education ecosystem informs ideas as to how to ‘connect the dots’ (Malin & Brown, 2020); for example Lysenko et al. (2016, p. 36) suggest that emphasis should be placed on the systemic nature of research knowledge utilisation and the multiplicity of agents involved, in which the process of transforming research evidence into useful knowledge is the critical component of the system.

Evidence that teachers do not use research to inform their teaching because they lack the skills or motivation to do so (Biddle & Anderson, 1991) and evidence that researchers do not communicate research in ways teachers can understand and use (Cai et al., 2020), taken together, suggest that, to tackle the community dissonance problem, either:

• teachers become better at accessing, reading, using and critiquing research, and
• researchers become better at communicating their ideas and their implications to teachers

or

• knowledge brokers explicitly take on both these barriers in a specialised and expert manner, potentially through creating boundary objects, which we explore in more detail in the following section.

The transformation of knowledge from research into classroom practice involves ‘a mix of complex processes, some of which need specialist mediation and dedicated resources’ (Cordingley, 2008, p. 37). Proposed models of knowledge brokering often suggest, however, that the role of knowledge broker may include multiple strands of work and influence—not just the movement of research through creating these resources, but the brokering of new discourses and the facilitating of critical dialogue around them. For example, Meyer (2010, p. 120) noted that brokering knowledge ‘means far more than simply moving knowledge—it also means transforming knowledge’, and suggested three types of knowledge brokering activity: identification and localisation of knowledge; redistribution and dissemination of knowledge; and rescaling and transformation of this knowledge.

If the community dissonance between teachers and researchers is to be tackled, it is crucial to consider therefore how and why research is produced alongside how and why it is used—to create loops of dialogue, including feedback, between teachers and researchers, as well as other educational decision-makers. Lavis et al. (2003) referred to this as developing a decision-relevant culture among researchers and a research-attuned culture among decision-makers. This careful consideration of multiple perspectives on research creation and dissemination as well as use, we argue, begins to fulfil some of the functions of Goldacre’s (2013, p. 15) information architecture system in that it creates the conditions for
building ‘bridges joining the people and strands of work together’. It has been hypothesised that lowering levels of community dissonance will lead to greater and deeper use of research (Farley-Ripple et al., 2018).

**Boundary objects**

Consideration of the community dissonance between education researchers and classroom practitioners—in particular the discourses of the two communities—have prompted some researchers to consider the idea of a *boundary object*. This is defined as an artefact that can ‘potentially be used to bring the activities of different communities into conversation with one another, and thus prepare the ground for “boundary crossing” moves’ (Venkat, 2017, p. 183). Moves to bridge the gap between teaching practice and research by creating such boundary objects are ‘likely to require judicious selections and adaptations of artefacts’ (ibid., p. 186). Star and Griesemer (1989, p. 393), the originators of the term ‘boundary objects’, describe them as ‘objects that both inhabit several intersecting worlds and satisfy the informational requirements of each of them…[remaining] robust enough to maintain a common identity across sites’. In other areas, such as product development, the theory of boundary objects has been used to share information successfully, for example by applying a pragmatic view of knowledge as localised, embedded and invested in practice (Carlile, 2002). It has been suggested that products of knowledge brokering, in particular summaries of research for users, may act as boundary objects in important ways. For example, one study examined the creation of scientific news stories as boundary objects:

> The properties of the news stories that enable them to act as boundary objects include their mutability and shareability as electronic draft documents created in distributed learning environments and responded to by an editor based at a distance, their flexibility in terms of genres for expressing the meaning and importance of science, and their portability across space in the form of print- and web-based publications that are valued by multiple communities. (Polman & Hope, 2014, p. 341)

Using boundary objects can be meaningful and useful because they are able to span the worlds of research and practice and speak to both groups of interest to our study, namely, researchers and practitioners in education. Also, boundary objects can facilitate and enable dialogue between communities, and hence one application of boundary objects could be as part of a knowledge brokering activity to address the community dissonance between researchers and practitioners in education (Farley-Ripple & Grajeda, 2020).

**The nature of evidence**

So far, we have considered the ‘how’ of knowledge brokering, but exactly what constitutes research evidence—the ‘what’ that is brokered from research into practice—is contested for several reasons. It is clear that a vital function of knowledge brokering is to select which research to broker (e.g. Meyer, 2010), but it is not yet clear how best to do this in a way that is transparent, ethical and accountable. There is a tension between the ‘scientific’ vision of reporting research results and the interpretative, storytelling process of knowledge brokering, likely to be unreplicable, that Hubers and Poortman (2018) suggest is importantly the broker’s unique vision.
In broad terms, fallibilism asserts that there is no certain or secure knowledge (Swann & Pratt, 2004) and postpositivism that no research is neutral or assumption-free (Phillips & Burbules, 2000) suggesting educational research findings ought never to be communicated (nor received) uncritically. Neither research nor teaching are apolitical, and conceptualising the movement of ‘useful’ research into the classroom as neutral or non-ideological is disingenuous (Collins, 2004, p. 69). There is a tension here between brokering research in a clear, useful manner, and smoothing out nuance and knots which may serve to faithfully signpost the interesting texture of research evidence to teachers. Equally,

there is a widespread and time-honoured tradition of privileging knowledge generated in academia over evidence coming from practical contexts, which deepens the gap between research and practice. Professional judgement based on the practitioners’ ‘soft’ data or tacit knowledge is often denied a place in actionable knowledge, consequently narrowing the practitioners’ role to only that of a consumer of externally produced knowledge. (Lysenko et al., 2014, p. 3)

Relatedly, it has been suggested that the choice of what kinds of knowledge to broker may have far-reaching implications in terms of privileging or devaluing types of knowledges, rather than integrating them (Swann & Pratt, 2004). This may apply across research evidence and other types of knowledge (e.g. teacher knowledge, community knowledge), or within research paradigms. Cai et al. (2020) found that a significant problem not just for teachers but for mathematics education researchers when considering a teacher viewpoint was to understand which types of evidence are relevant and useful compared to which have been more privileged in educational discourse.

Put together, the important considerations of what constitutes evidence in educational settings suggest specialists (referred to as knowledge brokers) who understand these—sometimes conflicting—practices and priorities of both teachers and researchers may be best placed to engage in intentional knowledge brokering in education (Cordingley, 1999).

**Research summaries**

Cordingley (2008) proposed that transformation of knowledge from research into classroom practice involves the creation of particular, specialist resources. As all too often education research is presented ‘in a form or medium which is largely inaccessible to a non-academic audience and lack[s] interpretation for a policy-making or practitioner audience’ (Hillage et al., 1998, p. xi), the creation of a type of mediating artefact which may act as a boundary object is suggested as part of this transformative process.

All research is subject to interpretation at various stages in production, and this can create problems in many ways: it may lack fidelity to the evidence; it may be repetitive; it may add no new understanding; it may neglect context or other important aspects; it can be useless, or simply boring (Smeyers et al., 2014). If ‘the basic unit of knowledge translation...[is] up-to-date systematic reviews or other syntheses of research findings’ (Grimshaw et al., 2012, p. 1), then it is implied that research summaries should seek to be up-to-date, systematic and/or narrative. Also, a significant amount of trust should be placed in the writer/interpreter. Yet Shank and Brown (2007, p. 19) advise the reader to exercise caution when reading secondary research articles—and ultimately to go back to the original research when unsure—because ‘writers differ in their abilities to interpret and summarize...and to understand and to draw conclusions and implications from their understandings’. This last point suggests more of the community dissonance between teachers and researchers—teachers are time poor (See et al., 2016) and often lack confidence or skills to make critical judgements on
WHAT MAKES A GOOD EDUCATIONAL RESEARCH SUMMARY?

It is likely that all research summaries will be biased to some extent, because the horizons (made from experience, scholarship, rigorous training and understandings) brought to research can also act as blind spots in a research synthesis (Smeyers et al., 2014); therefore the authors of research summaries should seek to minimise and/or render transparent this bias.

‘Research synthesis’ is in itself a part of research, a form of ‘doing’ science, and has been likened to bricklaying, where theorists provide blueprints and researchers collect the data to form those bricks, which are many sizes and shapes (H. Cooper & Hedges, 2009). Literature reviews, which are the result of considering research and ensuring it is ‘digested, sifted, classified, simplified, and synthesized’ (Manten, 1973 in Cooper & Hedges, 2009, p. 109), may have many purposes, only one of which is a summary of practices or applications for practitioners. For those research summaries which do share this purpose, some success criteria may be universal; Phipps et al. (2020) found that clear language, usability and good design (although the meaning of this is not completely clear) were shared features of all successful research summaries.

Shank and Brown (2007, p. 17) categorised secondary articles as ‘the work of someone who has read primary research articles or reports and then tells us their thoughts and understandings...not usually based directly in research findings, but on interpretations [of them]’. Of these, they identify the ‘lay review’, the ‘focused review’, the ‘interview’ and the ‘opinion piece’ as sub-categories. What we are referring to as a ‘research summary’ best fits with their subcategory of ‘focused review’: ‘what research can tell us about issue X’, where the writer serves as ‘mediator’ for those who do not have the time or skills to seek out all the original literature (ibid., p. 18).

There is still a lack of strong empirical evidence on the use of research summaries to improve practice in education (Farley-Ripple & Grajeda, 2020; Sharples, 2015). This is partly because the boundary between policy and practice is blurred (Hubers & Poortman, 2018), and also because conceptualising and hence measuring use of research in this way is complex and the field of knowledge brokering is still young (Rycroft-Smith, this issue). In the meantime, studies from outside the educational arena and often focused on policy may shed some limited light. For example, in healthcare, one study found that research summaries make research easier to understand, but do not necessarily increase the use of evidence from systematic reviews in policy making (Petkovic et al., 2018); and in international development, researchers found that the research summary or policy brief (presuming they are understood to be interchangeable) had a fairly weak effect on readers’ beliefs (Beynon et al., 2012).

**Summary**

It is suggested that using educational research to improve educational practice, in particular through knowledge brokering activity, is generally desirable. There are barriers and conflicts between the practices of teachers and researchers, collectively known as community dissonance, that should inform the design of such brokering activity. Research summaries for teachers could be an important part of this process, perhaps by serving as boundary objects. To help challenge dissonances and support brokering work beyond just the creation of resources, and make the knowledge flow effectively as well as being ethical, research summaries should be as successful as possible at communicating the research to teachers without disempowering them, perhaps by being intentional and transparent about choices around research selection and interpretation. To ensure research summaries act as boundary objects and do not imbalance relationships between teachers and researchers, the field should deepen its understanding of what teachers and researchers consider
effective practice in communicating the research through research summaries, and investigate areas of consensus and difference. Relating the characteristics of effective research summaries to their purposes and possible uses, which may be varied, can help enhance this understanding.

**METHODS**

**Study overview**

The method of comparative judgement was used to ask a group of 28 mathematics teachers and a group of 19 mathematics education researchers to rank a series of research summaries designed for teachers, from different sources. Follow-up questionnaires were given to all participants to ask about features of research summaries, and the answers given were grouped into themes and analysed.

**Comparative judgement (CJ)**

Comparative judgement emerged relatively recently in the field of educational assessment as a ‘viable, valid and highly reliable alternative to traditional marking’ (Tarricone & Newhouse, 2016, p. 251) and rests on the assumption that ‘all judgments are comparisons of one thing with another’ (Laming, 2004, in Pollitt, 2012, p. 285). In comparative judgement for assessment of student work, ‘professional judgement by teachers replaces the marking of tests; a judge is asked to compare the work of two students and simply to decide which of them is the better’ (Pollitt, 2012, p. 281). Specifically,

> For each pairing of test responses, an assessor must decide which has the greater of a specified global construct...The outcome is a binary decision matrix of the ‘winner’ and ‘loser’ for each pairing. The matrix is then fitted to a statistical model, such as the Bradley–Terry model to produce a parameter estimate [which] can be used to construct a scaled rank order of test responses from ‘best’ to ‘worst’. (McMahon & Jones, 2015, p. 371)

In this study we sought to create a ranking from a sample of research summaries in each group of participants (mathematics teachers and mathematics education researchers) and then compare them, for which comparative judgement was a judged good fit as a way of measuring consensus both within a group and across two groups, and without having to provide a particular set of criteria on which to judge. Providing a set of criteria would have been problematic because the second research question focused on emergent features of research summaries that participants themselves considered important. Indeed, ‘CJ uses no detailed assessment criteria or scoring rubrics and the final rank order is instead grounded in the collective expertise of the judges’ (Bisson et al., 2016, p. 143). Bisson et al. also suggest that:

> The shift from rubrics to a reliance on collective expertise for measuring understanding can be an uncomfortable notion, and is sometimes viewed by those used to traditional measurement methods as opaque and under-defined. However we argue that this is a key strength: a given concept is defined by how it is perceived, understood and used by the relevant community of expert practitioners.
It has been suggested CJ can be used therefore to evaluate items that are considered important but resist criterion-based assessment (B. Davies et al., 2021). CJ is still a relatively new method to emerge in the literature, and had not been used before in the particular way we used it in this study. However, in many ways the first research question is analogous to a group of teachers wishing to grade student work that is more complex than a straightforward performative answer to a closed question, that is, ‘a sustained performance that reflects the discourse and cultural norms of the subject’ (Jones, Swan, et al., 2015, p. 152). Indeed, CJ has been used successfully to assess portfolios, narrative writing and mathematical problem solving (Jones & Inglis, 2015). The use of CJ continues to grow in more recent research, for example with students in product design to support their understanding of quality (Ivo et al., 2021) and with mathematicians to investigate their conceptions of the notion of proof (B. Davies et al., 2021). Recently, Sangwin and Kinnear (2021) also used CJ to judge several different aspects of mathematical proof, evaluating whether they were reliable constructs by post-hoc analysis.

On the whole, CJ is considered effective in evaluation of complex, creative and composite work involving skills such as making connections between learned ideas, applying understanding to novel contexts, constructing arguments and demonstrating chains of reasoning (Marshall et al., 2020) which we suggest includes knowledge brokering work such as the research summaries evaluated in this study.

Finally, CJ relies on asking as broad and simple a question as possible to allow a holistic judgement to take place (Pollitt & Whitehouse, 2012). The question ‘Which of these communicates the research better to teachers?’ was therefore chosen to be displayed when participants were shown each time a pair of research summaries to choose between.

This study used pairwise CJ, in which the judges were given pairs of research summaries (shown on screen as two separately scrollable and zoomable PDFs) and was conducted using the Cambridge Assessment custom-built CJ platform.¹

### Procedure

The data were collected in three parts:

1. A screening questionnaire was given to separate participants into either the mathematics teacher group or the mathematics education research group, and to collect some background information about them.
2. Each participant was given a CJ pairwise task to complete in their own time and individually, in which each judge saw between 7 and 11 pairs of research summaries.
3. Each participant was sent a follow-up questionnaire, with screenshots of their most and least highly rated research summaries to help prompt reflection.

The questions asked in part 3 were:

- You suggested this research summary (attached to email; snippet shown here) may have communicated the research well to teachers. Which features or factors do you think may have contributed to this, or did you find elsewhere in the study to be part of your decision making?
- You suggested this research summary (attached to email; snippet shown here) may have communicated the research less well to teachers. Which features or factors do you think may have contributed to this, or did you find elsewhere in the study to be part of your decision making?
- Do you have anything further to add about the study or your experience of completing the task?
Pilot

A pilot study with 20 research summaries was conducted with nine teachers who were studying the Master's course in mathematics education at the University of Cambridge, and one mathematics education researcher. The groups were not separated in this case as the second group would have had only one participant. Also the purposes of the pilot were consistent with one large group. The purposes were:

- To observe the participants’ reactions to being asked to complete a comparative judgement task;
- To consider the effect on participants of the number of pairwise judgements (and therefore the total number of research summaries in the study);
- To observe and plan for any technical issues;
- To collect any other written feedback;
- To collect initial summary empirical data for the pilot group, placing it within the context of the participants in the pilot, in particular reliability measures;
- To use time taken to make judgements as a guide for participants in the final study.

Running the pilot study suggested the following design elements in the final study:

1. Participants reported comparing 16 or 17 research summaries in total as ‘overwhelming’ and ‘too many’—therefore the total number of pairwise comparisons was reduced for the final study to between 7 and 11 (the discrepancy here is due to the fact that the two groups had different numbers of participants and the number of comparisons per research summary was kept constant).
2. The length of documents was raised as an issue, with some participants questioning whether comparing a two-page document and a 20-page document was reasonable. After some consideration, we decided not to eliminate any research summaries from the final study based on length, as we wanted to see whether length would emerge as an element in participants’ own analysis of the features involved.
3. The ability to leave the task and return to it later was cited as important; this was something we therefore emphasised in the instructions to participants in the final study.
4. Some participants were initially unsure of the criteria for judging, asking how much scope there was for their own criteria of importance; this was something we again made clear in the instructions to the participants for the final study.
5. There were some technical issues such as loading time and use of different browsers/devices affecting ability to see all research summaries clearly; we therefore made it clear in communication with participants in the final study that technical support would be provided where needed.
6. The Scale Separation Reliability (SSR) score for the pilot was 0.72, suggesting a good level of consensus as to more and less successful research summaries among participants. SSR is a measure of reliability described as ‘the ratio of true variance to observed variance in the estimated scale values’ (Holmes et al., 2017, p. 20) where 0 indicates the lowest reliability and 1 the highest.
7. Several participants reported enjoyment of the method of comparative judgement, stating that the process of comparing was both simple to carry out and interesting, which suggested this might be a view shared by participants in the final study.
WHAT MAKES A GOOD EDUCATIONAL RESEARCH SUMMARY?

Choice of research summaries

To address our research questions, it was crucial that the items chosen as representing mathematics education research summaries (a) fitted the description synthesised from the literature review, and (b) were as diverse as possible within constraints of time and language requirements. The latter is a recommended element when curating items for comparative judgement studies (McMahon & Jones, 2015). We researched different organisations and documents that stated as their purpose summarising, interpreting, or commenting on research for (mathematics) teachers, limiting our search to English-speaking countries, and found ten different ‘libraries’ of such research summaries, of which we chose two examples from each.

One of these libraries contained items which one of us (first author) had authored. Therefore, it was particularly important for ethical and sampling considerations that we did not personally choose the specific artefacts from each library but used a transparent random process to do so. We did this by counting the total elements publicly available (without cost or membership) in each library, ordering them by date, and using a random number generator to select two of them. This process meant that some of the research summaries were less focused on mathematics education research than others, as some of the libraries were more broadly applicable to education and some more focused on mathematics specifically. It also meant that some of the research summaries were more recently published than others. The range of dates of publication, along with other details, can be seen in Table 1. For use in the CJ task, each article was either downloaded as or converted to pdf format.

Recruitment of participants

Since the goals of this research were to harness the collective expertise of two groups of judges in a CJ task, it was important to the validity of this study to have a reasonable number of participants in both groups. As such, a ‘convenience sample’ from a specialised population was used; participants were recruited to the study via a combination of social media posts (Twitter and Facebook posts in mathematics teacher groups) and direct emails (mostly to mathematics education researchers). This is likely to have resulted in a biased sample thus impacting generalisability; for example, teachers who use Twitter are generally more research-engaged (Carpenter & Krutka, 2015) and recruitment through personal connection is likely to result in a less diverse sample (Watson & Moritz, 2000). However, the use of a screening questionnaire as the first part of the study allowed us to collect some data on the participants in order to examine these issues in more detail.

Of a total of 35 teachers who completed the screening questionnaire, 28 completed the study. Three were primary school teachers, one was a middle school (Years 5–7) teacher, and 24 were secondary teachers. Figures 1 and 2 show the self-reported research literacy of participants in both groups as measured by a question in the screening questionnaire which asked ‘How familiar are you with reading, discussing and critiquing education research?’ which aligns with the definition of educational research literacy for teachers used by Shank and Brown (2007). Research literacy was fairly balanced across the teacher participants, as shown in Figure 1. Figure 2 shows the self-reported research literacy of participants in the mathematics education researcher group, which by contrast (and as expected) is very skewed. Figures 3 and 4 show the years of experience in mathematics education of the two sample groups.
**TABLE 1** The research summaries chosen for the study

<table>
<thead>
<tr>
<th>Organisation</th>
<th>Library/journal</th>
<th>Stated purpose</th>
<th>Two chosen artefacts (and their length in pages)</th>
<th>Dates of publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cambridge Mathematics</td>
<td><em>Espressos</em></td>
<td>‘a small but intense draught of filtered research on mathematics education, expressly designed with teachers in mind’. <a href="https://www.cambridge.maths.org/espresso/">https://www.cambridge.maths.org/espresso/</a></td>
<td>Espresso 15: Introducing negative numbers (2)</td>
<td>Dec 2018</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Espresso 10: Working memory for mathematics learning (2)</td>
<td>Dec 2017</td>
</tr>
<tr>
<td>The Australian Association of Mathematics Teachers</td>
<td><em>Australian Primary Mathematics Classroom/ The Australian Mathematics Teacher</em></td>
<td>‘directed toward mathematics educators… reports and ideas for teaching activities, as well as reporting research which has a direct impact on classroom teaching’. <a href="https://aamt.edu.au/teachers/journals/aamt-journals/">https://aamt.edu.au/teachers/journals/aamt-journals/</a></td>
<td>The ‘art’ of ratio: Using a hands-on ratio task to explore student thinking (8)</td>
<td>March 2017</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Reflecting on fraction addition: Using the mirror formula to visualise unit fraction addition (6)</td>
<td>June 2018</td>
</tr>
<tr>
<td>Association of Teachers of Mathematics</td>
<td><em>Mathematics Teaching</em></td>
<td>‘articles which reflect on the practice of teaching mathematics… a balance between those derived from research and from practical experience… clarity, topicality, the extent to which they reflect upon knowledge and understanding of mathematics teaching and learning, and their contribution to inspiring further development and research’. <a href="https://www.atm.org.uk/Mathematics-Teaching-Journal-Archive">https://www.atm.org.uk/Mathematics-Teaching-Journal-Archive</a></td>
<td>Symbol non-sense: A cautionary tale (4)</td>
<td>July 2018</td>
</tr>
<tr>
<td>The Mathematical Association</td>
<td><em>Mathematics in School</em></td>
<td>‘aimed at teachers of secondary pupils, students in training and all those with a professional interest in mathematics education… contributions from mathematics practitioners at all levels and articles that reflect the best of current thinking and practice by means of exploring further development and sharing practical experiences’. <a href="https://www.m-a.org.uk/mathematics-in-school">https://www.m-a.org.uk/mathe matics-in-school</a></td>
<td>How is Mathematical Language Used in the Classroom? (6)</td>
<td>Jan 2013</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Working with trainee teachers on their use of digital technologies in the classroom (5)</td>
<td>May 2014</td>
</tr>
<tr>
<td>Harvard Graduate School of Education</td>
<td><em>Usable knowledge: research stories</em></td>
<td>‘Relevant research for today’s educators… to connect research to practice. We make education research and well-vetted strategies accessible to a wide audience: teachers and principals, district leaders, policymakers, university faculty and higher ed professionals, nonprofit leaders, entrepreneurs, members of the media, and parents… rigorously reported, credible, and useful content’. <a href="https://www.gse.harvard.edu/about">https://www.gse.harvard.edu/about</a></td>
<td>Confronting gender bias at school (3)</td>
<td>Sept 2015</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tests and stress bias (3)</td>
<td>Feb 2019</td>
</tr>
</tbody>
</table>
### Table 1 (Continued)

<table>
<thead>
<tr>
<th>Organisation</th>
<th>Library/journal</th>
<th>Stated purpose</th>
<th>Two chosen artefacts (and their length in pages)</th>
<th>Dates of publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mapping Educational Specialist KnowHow (MESH)</td>
<td>MESH guides</td>
<td>‘We provide teachers and other educators with quick access to summaries of research-based specialist knowledge to support their professional judgement…research summaries, updated periodically, written for teachers (and parents/carers), linked to the research giving rise to the advice’. <a href="http://www.meshguides.org/what-is-mesh/">http://www.meshguides.org/what-is-mesh/</a></td>
<td>Reading and writing mathematics (1) Dyscalculia: what the beginning teacher needs to know (14)</td>
<td>March 2013 (revised Mar 2016) May 2019</td>
</tr>
<tr>
<td>Education Endowment Foundation</td>
<td>Teaching and Learning Toolkit</td>
<td>‘An accessible summary of the international evidence on teaching 5–16-year-olds… Accessible summaries of educational research to guide teachers and senior leaders on how to use your resources to improve learning outcomes’. <a href="https://educationendowmentfoundation.org.uk/education-evidence">https://educationendowmentfoundation.org.uk/education-evidence</a></td>
<td>Social and emotional learning (6) Setting or streaming (10)</td>
<td>Feb 2019 Nov 2018</td>
</tr>
<tr>
<td>National Centre of Excellence for the Teaching of Mathematics</td>
<td>Evidence bulletins</td>
<td>‘Articles about research into mathematics education… The evidence bulletin is designed to summarise research in interesting ways at the same time as providing materials to support teachers who want to try out new methods’. <a href="https://www.ncetm.org.uk/media/vzgheehz/issue-2-primary-magazine.pdf">https://www.ncetm.org.uk/media/vzgheehz/issue-2-primary-magazine.pdf</a></td>
<td>What makes some teachers of mathematics more effective than others? (1) How can whiteboards help to make learning interactive? (1)</td>
<td>May 2009 May 2009</td>
</tr>
<tr>
<td>Institute for Effective Education (IEE) and the Center for Research and Reform in Education at Johns Hopkins University</td>
<td>Best Evidence in Brief fortnightly newsletters</td>
<td>‘a round-up of the latest education news… we focus on stories with practical implications for schools and policy makers, and try to include only high quality research’. <a href="https://the-ieee.org.uk/what-we-do/best-evidence-in-brief/">https://the-ieee.org.uk/what-we-do/best-evidence-in-brief/</a></td>
<td>Best Evidence in Brief 16 Oct 2008 (6) Best Evidence in Brief 8th May 2012 (1)</td>
<td>Oct 2008 May 2012</td>
</tr>
<tr>
<td>Institute of Education Sciences, US Department of Education</td>
<td>What Works Clearinghouse intervention snapshots</td>
<td>‘We review the research, determine which studies meet rigorous standards, and summarize the findings. We focus on high-quality research to answer the question “what works in education?”’ <a href="https://ies.ed.gov/ncee/wwc/WhatWeDo">https://ies.ed.gov/ncee/wwc/WhatWeDo</a></td>
<td>Fraction Face-Off! (1) Pre-K mathematics (3)</td>
<td>March 2020 Dec 2013</td>
</tr>
</tbody>
</table>
The first measure used to analyse the CJ data was the Rasch scale separation reliability (SSR) which lay underneath each group's rank order; this is a measure considered by some to be analogous to Cronbach’s α (e.g., Wright & Masters, 1982). We also utilised an exploratory data analysis (EDA) approach (e.g., Tukey, 1977), considering the ranking versus the ‘measure’ statistic calculated by the Bradley-Terry model; comparing the two rankings by eye to look for patterns, clumps and clusters; analysing the ways in which the two rankings produced by the two groups (teachers and researchers) correlated, using Spearman's rank to calculate basic correlation and analysing visual features in more detail; and using a Sankey-style diagrammatic analysis to consider ‘flow’ between rankings of the two groups.

A thematic analysis approach was used to analyse the data collected from the three follow-up questions, which 82% (23/28) of teachers and 89% (17/19) of researchers who completed the CJ task responded to. Thematic analysis in qualitative research may involve: discovering themes and subthemes; winnowing themes; building hierarchies or structures of themes; and linking themes into theoretical models (Ryan & Bernard, 2003). Discovering themes involves looking for repetition, specialist language, metaphors and analogies, similarities and differences (ibid.); and considering deep structure by seeing the process as
WHAT MAKES A GOOD EDUCATIONAL RESEARCH SUMMARY?

jazz-like (Oldfather & West, 1994). We used an inductive approach to analysis, allowing the theory to emerge from the data (Barnett-Page & Thomas, 2009). The results of all three questions answered by teachers and, separately, researchers, were analysed in this way, allowing themes that were the same and different to emerge across the two groups. The unit of analysis was each different point made by participants in their question responses, that is, it was applied flexibly rather than strictly at sentence, paragraph or response level.

We analysed all three questions together after noticing that some respondents freely mixed in positive and negative features in all three questions, or misread questions of positivity to suggest negativity and vice versa. We used a mixed approach: categories of notions with similar meaning and connotations were developed, relying on both a priori and a posteriori approaches (Leech & Onwuegbuzie, 2011), specifically using elements of constant comparison analysis, classical content analysis, domain analysis, taxonomic analysis and componential analysis.

**Reliability**

Comparative judgement ‘achieves extremely high levels of reliability, often considerably higher than practicable operational marking can achieve’ (Pollitt, 2012, p. 281). Yet, more
recently, Verhavert et al. (2019) in a meta-review of CJ found a large variability in reliability between assessments, both in the way they are implemented (e.g., type of assessed performance or number of comparisons) and in the reliability of the results. Reliability in CJ methodologies is typically measured by comparing to traditional assessment, or as in this study, estimation of the scale separation of the outcomes using scale separation reliability (SSR). Verhavert et al. (2018, p. 444) found that ‘there are good and strong indications that the SSR reflects the interrater correlations [in CJ] but some results call for caution’. A significant criticism has been made that ‘adaptive’ CJ (using nonrandom pairings) inflates reliability scores (Verhavert et al., 2018), hence the non-adaptive version was used in this study.

The choices we had to make for the study design in terms of reliability were:

1. number of research summaries
2. number of judges per group
3. number of judgements per research summary
4. number of judgements per judge

The first of these variables is driven by the scope of the project; we aimed to recruit 30 in the teacher group and 20 in the researcher group, and achieved 28 and 19 respectively. We also chose a total of 20 research summaries from ten different libraries as a number that seemed manageable yet allowed for a good range of variation, an important factor in CJ design. These being fixed variables, there was a tension between the remaining two elements when considering reliability, because after the pilot we were cautious that judges not be overwhelmed with too many judgements; yet a certain threshold number of judgements per artefacts was important to achieve. We discussed these with an expert who emphasised...
that it was important to run a pilot, with two main considerations the most important: the number of judgements (ideally more than ten per artefact) and the qualitative difference between artefacts (the more of a range, the better). This consultation also suggested a reliability rating as measured by SSR of between 0.5–0.6 would be sufficient for a small-scale study such as this for which the focus was not necessarily just on being reliable, but considering reliability as part of the research questions.

The SSR score for the pilot ($n = 10$) was 0.72, suggesting a good level of consensus as to more and less successful artefacts among participants; however, it has been found that ‘SSR declines as fewer and fewer judgements are made for each study’ (Verhavert et al., 2018, p. 440) and as such we would not expect a high (>0.9) SSR for a small-scale study with two groups such as this. In a meta-analysis of comparative judgement research, Verhavert et al. (2019, p. 2) found that ‘for reliability levels of .70 between 10 and 14 comparisons per performance are needed. This rises to 26 to 37 comparisons for a reliability of .90’, which was a concern after pilot participants suggested 17–18 comparisons was overwhelming.

However, Holmes at al. (2017, p. 28) conducted an in-depth study of multiple judgements of the same artefacts, suggesting that the total judgement number was more of a focus for reliability: ‘there is a steady decline in SSR as the judgement number is reduced, which becomes increasingly steep and below about 150 judgements the model fit starts to fail as indicated by the variable SSR values’. For this reason, we designed the study to ensure each of the 20 research summaries was judged 20 times, giving a total of 200 judgements for each of the two groups (as each judgement involved two research summaries) rather than focusing on the number of judgements per judge. Holmes et al.’s (2017) data suggested, for marking student work, an SSR of around 0.7 would have been measured for 200 judgements, giving an approximate benchmark to measure reliability against.
This meant that for the teacher group ($n = 28$), judges only made seven or eight judgements in total, which we might also expect to be reflected in the SSR value. Verhavert et al. (2019) reported that, for low-stakes or formative assessment, a reliability level of 0.7 or greater is deemed sufficient, whereas for high-stakes or summative assessment, 0.9 is typically used as a threshold. As this was an exploratory study, using CJ in an innovative manner, we settled on an SSR target of approximately 0.7 to be suitable to judge a level of within-group consensus.

**Validity**

Validity is widely debated in the educational assessment literature (McMahon & Jones, 2015) but concerns the extent to which the judging procedure chosen uses the evidence to produce valid outcomes, or ‘whether what is being measured is what the researchers intended’ (Billington & Meadows, 2005, p. 13). The validity of results in this study is supported by a single, broad question in the CJ study and the inclusion of two separate and reasonably large groups of expert judges from different communities. In CJ, ‘validity is grounded in what is valued by the community of practice within a given discipline’ (Jones, Swan, et al., 2015, p. 153), and here we define two separate communities of practice—mathematics teachers and mathematics education researchers—with the aim of comparing the results to examine areas of consensus and difference. In CJ, validity and reliability are less in conflict than in traditional assessment, and can be balanced by a large number of judges with a good degree of diversity, as in this study, ‘providing a more comprehensive image of the competence’ (Verhavert et al., 2019, p. 30).

One reported threat to validity when using teacher assessment of student artefacts is teacher bias based on knowledge of students (Bennett et al., 1993). In this study the latter could translate as bias in the judges based on broader knowledge of or relationships with the authors, institutions or knowledge brokers of the research summaries, which information was clearly visible on the research summaries in the CJ task. As described earlier, we specifically did not choose to anonymise research summaries because considerations around credibility and trust were part of the emergent themes that might have arisen from the data in response to the research questions.

There is an interesting tension here around validity and reliability: if validity is defined as the extent to which a test measurement reflects the items being measured and not those doing the measurement (Billington & Meadows, 2005), then CJ as the methodology in this study is dependent on the emergent characteristics being (a) stable in some way and (b) representative of the thinking of the community; using different groups comparatively as in this study design seeks to analyse some of these issues. We are aware that we are not therefore using CJ in the way most previous studies have done so; to the best of our knowledge, this is the first time CJ has been used to assess teacher-facing research materials in this way. Pollitt (2012, p. 282) suggested that CJ is valid for assessing ‘performances, portfolios, reports and likely more’ and we expect to see more discussion of the meanings of validity and reliability in such contexts as CJ is used more broadly.

**Ethics**

The study was reviewed by the University of Cambridge Faculty of Education Research Ethics Committee and received ethics clearance. The first ethical consideration for the study design was concerns around the context of knowledge brokering that emerged from the literature review, which encompassed ideas around the nature of evidence, the existence of
community dissonance, and the importance of multiple perspectives. Trying to be balanced about the use of evidence meant that there was a possibility we would include research summaries which might not have presented evidence in the most balanced manner, might have privileged certain types of evidence, or might not have used rigorous or transparent sources of evidence. We made the decision to use research summaries from reputable organisations to mitigate these concerns as much as possible. At the same time, we acknowledge that we are not, and cannot be, the arbiters of such standards, and in fact this was in part the purpose of the study: to explore whether mathematics teachers and mathematics education researchers found any of these features to be important when judging research summaries for themselves. For this reason too, it was made clear in communications with participants that it was up to them to decide how to choose between the pairs of research summaries that were presented to them each time during the CJ task when considering which of the two research summaries communicated the research better to teachers.

Ethical research guidelines from BERA suggest ‘[s]ensitivity and attentiveness towards such structural issues [of power] are important aspects of researchers’ responsibilities to participants at all stages of research, including reporting and publication’ (BERA, 2018, p. 86). As discussed in the literature review section, there is some evidence that relationships between teachers and researchers may have been obstructed by ideas of hierarchy and structural barriers, influencing both knowledge and power flow. We were careful in our communication with the participants not to treat the two groups differently, sending exactly the same guidelines and text to both, with explicit inclusion of instructions which empowered participants to choose their own criteria when making decisions. This read:

How do you decide which is best? The question is simply ‘which of these communicates the research better to teachers?’ and you can decide that any way you like—read them in detail, scan them, get a ‘feel’ for which you think is more successful, but focus on ‘at communicating the research to teachers’. There is no ‘right’ way to answer the question.

Another important ethical aspect of this study which relates to power and relationships is the fact that, as we noted earlier, one of us (first author) was the author of two of the research summaries. We decided to leave all the author names, as well as the branding, on the research summaries partly because we wanted to see if issues of trust and credibility emerged from the results, and hoped that, if participants felt any kind of pull to rank higher those research summaries that one of us had authored, they would mention it when asked the open-ended question at the end (‘Do you have anything further to add about the study or your experience of completing the task?’). Since we were exploring issues around features of successful research summaries, removal of information that related to trust and credibility could be seen to be unnecessarily introducing bias at the first stage of the study. Similarly, because education researchers who may identify as knowledge brokers were participating in the study, there was a chance that participants could also be authors. This in fact did happen in at least one case, although this participant did not see the research summary they had authored because the random pairings did not match them with that particular item. Yet another reason for our choice to leave all author information on the research summaries was our belief that this accorded with one of Denscombe’s (2007) principles when analysing ethical considerations: researchers should avoid deception or misrepresentation. We gave our full names in communication, leaving participants informed that one of us was the author, but without explicit discussion of it in case this heavy-handed approach encouraged more bias than it prevented. Overall, we saw this as an ‘ethical dilemma’ as explored by Stutchbury and Fox (2009)—researchers cannot ‘solve’ this kind of dilemma by definition, but they can recognise, anticipate and discuss them, as we have here.
Finally, the principle of informed and continuous consent was adopted throughout the study (e.g. Denscombe, 2007); in the invitation letter, participants were told that ‘Throughout the research participants will be able to refuse to participate at any time.’ Participants’ information was collected and only kept for the stated purpose, in compliance with General Data Protection Regulations; participants were clearly informed of this and that their data would be used anonymously. Participants were also clear that they could email at any time to provide feedback or clarification, or ask for support on the process, and several did, to which we responded promptly wherever possible. All questions in the follow-up questionnaire were open-ended with no limit on length, and the final question was completely open-ended, allowing participants a voice to communicate critically with us about the study if they wished.

Limitations

A limitation of this study is the need for expertise in judges, as CJ has been criticised as subjective; ‘it is essential that CJ is conducted by many experts so that biases are cancelled out’ (Bisson et al., 2016, p. 143), a consideration reflected in our research design. Another limitation is the extent to which judging between the research summaries used here is indeed comparing like for like (artefacts aiming to fulfil the same purpose). Yet another limitation is the likely variations in time taken, effort levels, and levels of distraction between participants, especially as previous studies have found that some teachers show patterns of interruption when completing CJ studies. However, evidence suggests that judges who spend longer making decisions are no more accurate than those who complete their judgements quickly (McMahon & Jones, 2015). Finally, the study is limited by the sample size and characteristics, which suggests the results are exploratory only and should not be generalised without caution to the general population of mathematics teachers and mathematics education researchers.

FINDINGS

Comparative judgement findings

The findings presented in this section address the first research question, pertaining to consensus within and across the two groups. Table 2 shows the CJ results for the mathematics teacher group. The ‘measure’ statistic is the parameter estimated for each research summary after fitting the Bradley-Terry model for pairs: it is a scaled estimate representing the judged quality of each research summary relative to the others, meaning artefacts with a higher measure were better (i.e. won more comparisons). The SSR (within-group agreement) level of the mathematics teacher group was 0.61, indicating a moderate degree of reliability although not quite reaching the 0.7 target that was set in the study design. As we discussed in the study design, the different sizes of the two groups meant that if we wanted to keep the number of times each research summary was seen constant between groups, there was a consequential lowering of the number of judgements each judge made in the teacher group (seven or eight per judge) which is the likely reason for this slightly lower SSR.

The SSR level of the mathematics education researcher group was 0.67, once again indicating a moderate degree of reliability and only narrowly missing the 0.7 target that was set in the study design. This was likely due to the number of judgements each judge made in this group (10 or 11) being higher than the teacher group (as we elected to keep the number of judgements per research summary the same between groups), but it may also have been due to this being a more homogenous or a smaller group. The CJ results for the mathematics education researcher group are shown in Table 3.
WHAT MAKES A GOOD EDUCATIONAL RESEARCH SUMMARY?

The Spearman's rank correlation coefficient (i.e., the nonparametric version of the Pearson product-moment correlation) between the rankings of the two groups was measured as 0.43, indicating a moderate correlation. The p-value for the Spearman's rank correlation coefficient for these data was calculated at 0.059, narrowly failing the significance test at the 5% significance level. The correlation is therefore found to be moderate at the 10% significance level, which suggests results should be interpreted with caution. Given the clustering of the data and the importance of not placing too much weight on individual rankings created from this kind of pairwise comparison, this represents a tentative degree of consensus and begins to answer our first research question: Overall, teachers and researchers generally agreed, although only moderately, on which research summaries were more or less successful at communicating the research to teachers, with some research summaries clearly the exception to this (see Figure 5).

<table>
<thead>
<tr>
<th>Table 2 Results of the CJ task, mathematics teacher group, ( n = 28 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>18</td>
</tr>
<tr>
<td>19</td>
</tr>
<tr>
<td>20</td>
</tr>
</tbody>
</table>
We also plotted the paired ranking data as a dynamic ‘movement flow’ in a single flow Sankey-type diagram (see Figure 6). This allows a further visual comparison of individual rankings and clustering together, showing which research summaries had the least consensus (‘moved’ the most between the two rankings, evidenced by the steepest gradients) and which had the most consensus (‘moved’ the least between the two rankings). The graph shows more ‘tangling’ in the midsection. One research summary, the AAMT article *The art of ratio*, was ranked exactly the same by both teachers and researchers. The research summaries that caused the most disagreement between the two groups were the ATM article on the use of algebra in geometry *Symbol non-sense* and two Usable knowledge articles.

In Table 4, we have ranked the research summaries by the degree of consensus in their rankings between the two groups, measured by taking the difference between the position in the ranking given by the teacher group and the one given by the researcher group.
WHAT MAKES A GOOD EDUCATIONAL RESEARCH SUMMARY?

Questionnaire findings

All participants were sent a follow-up questionnaire after completing the CJ task, in which they were asked about features of the research summaries they had judged. The data were then analysed, keeping the two groups separate, to discover emergent themes. Each participant's questionnaire responses were divided into individual ideas expressed, called ‘items’ here for the sake of clarity (meaning that there were many more items than participants, as some people included several items in a single answer to a question on the questionnaire). The chosen themes and examples of items which were considered to be expressions of that theme are shown in Table 5; italics indicate text from the teacher group and no italics indicate text from the researcher group.

Table 6 shows the total proportion of each theme in the mathematics teacher group by considering the number of items that referred to that theme out of the total; these are expressed as percentages for ease of comparison.

As shown in Table 6, mathematics teachers considered graphic design to be the most important feature (as measured by the most mentions) in terms of a research summary which communicated the research well to teachers. They also considered being summarised or focused, having takeaways or implications for practice, being easy to read or accessible and having a successful structure to be important. The length of the research summary or the time it might take to read, the language used and whether the summary was written on an interesting or relevant topic were also mentioned.

For the mathematics education researcher group, as shown in Table 7, graphic design was also the most mentioned theme, although it was mentioned less as measured by the
proportion of items (12.5% compared to 20.6% in the teacher group). Mathematics education researchers considered the concept of being summarised or focused important in a research summary important, as well as the research summary being easy to read or accessible, and having takeaways or implications for practice as well as some attention paid to clear methodology or good use of evidence. They also mentioned length or time, structure and the need for links or easily available further information. Mathematics education researchers considered trust/credibility and the ability to prompt reflection or critical discourse for teachers fairly important, as measured by frequency of mentions in response to the follow-up questions.
WHAT MAKES A GOOD EDUCATIONAL RESEARCH SUMMARY?

When plotted as paired data, the proportion of items in each theme appears moderately (linearly) correlated, and the Product Moment Correlation Coefficient (PMCC) is measured at 0.49, further supporting this view of moderate agreement, as shown in Figure 7. However, the \( p \)-value is calculated at \( p = 0.089 \), failing at a 5% significance level but passing at a 10% level. Since the PMCC was not the focus of this methodology, and was only included to add to the holistic picture following an EDA approach (Tukey, 1977), we note that it appears in keeping with the rest of the analysis in the study. The \( p \)-value suggests that the likelihood of achieving such a moderate correlation by chance is approximately 8.9% which, along with the imprecision of assigning thematic categories and the small sample size, gives reason for caution when interpreting these results. However, we consider the PMCC a way of ‘adding to the picture’ here rather than the single way to answer a research question, as one might use it as part of a complete hypothesis test.

**Overlapping themes**

As shown in Figure 8, the overlap in questionnaire items as grouped by our chosen themes is considerable: 83.1% of the total items from the mathematics teacher group, and 65.4% of the items from the mathematics education researcher group, were in overlapping thematic categories. Yet this does not necessarily mean consensus in terms of point of view brought to that theme, just that the element was mentioned by participants in that group. It is also

### Table 4

<table>
<thead>
<tr>
<th>Research summary</th>
<th>Teacher position</th>
<th>Researcher position</th>
<th>Absolute difference in ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAMT: the art of ratio</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>MA: mathematical language in the classroom</td>
<td>5</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Best evidence in brief: May 2012</td>
<td>12</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>MESH guide: reading and writing maths</td>
<td>17</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>Espresso 15: Introducing negative numbers</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>MA: digital technologies in the classroom</td>
<td>15</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>NCETM evidence bulletin: interactive whiteboards</td>
<td>14</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>NCETM evidence bulletin: effective maths teaching</td>
<td>6</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>What Works Clearinghouse: pre-K mathematics</td>
<td>20</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>EEF: social and emotional learning</td>
<td>13</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>Espresso 10: Working memory for mathematics learning</td>
<td>1</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>ATM: cognitive load theory</td>
<td>10</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>EEF: setting or streaming</td>
<td>16</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Best evidence in brief: Oct 2018</td>
<td>7</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>What Works Clearinghouse: fractions faceoff</td>
<td>8</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>AAMT: fraction addition</td>
<td>11</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>MESH guide: dyscalculia</td>
<td>2</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>Usable knowledge: gender bias</td>
<td>18</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Usable knowledge: test and stress bias</td>
<td>9</td>
<td>19</td>
<td>10</td>
</tr>
<tr>
<td>ATM: algebra in geometry</td>
<td>19</td>
<td>5</td>
<td>12</td>
</tr>
</tbody>
</table>

When plotted as paired data, the proportion of items in each theme appears moderately (linearly) correlated, and the Product Moment Correlation Coefficient (PMCC) is measured at 0.49, further supporting this view of moderate agreement, as shown in Figure 7. However, the \( p \)-value is calculated at \( p = 0.089 \), failing at a 5% significance level but passing at a 10% level. Since the PMCC was not the focus of this methodology, and was only included to add to the holistic picture following an EDA approach (Tukey, 1977), we note that it appears in keeping with the rest of the analysis in the study. The \( p \)-value suggests that the likelihood of achieving such a moderate correlation by chance is approximately 8.9% which, along with the imprecision of assigning thematic categories and the small sample size, gives reason for caution when interpreting these results. However, we consider the PMCC a way of ‘adding to the picture’ here rather than the single way to answer a research question, as one might use it as part of a complete hypothesis test.

**Overlapping themes**

As shown in Figure 8, the overlap in questionnaire items as grouped by our chosen themes is considerable: 83.1% of the total items from the mathematics teacher group, and 65.4% of the items from the mathematics education researcher group, were in overlapping thematic categories. Yet this does not necessarily mean consensus in terms of point of view brought to that theme, just that the element was mentioned by participants in that group. It is also
<table>
<thead>
<tr>
<th>Theme</th>
<th>Example item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphic design</td>
<td><strong>Clean layout and design</strong>&lt;br&gt;The visual layout is not to my liking. In my opinion, changes of color and font and font size all increase the cognitive load and decrease the readability</td>
</tr>
<tr>
<td>Summarised/focused</td>
<td><strong>The obvious summaries make it easy to scan and see if I am interested in reading the content.</strong>&lt;br&gt;[I could see] the headlines from the research</td>
</tr>
<tr>
<td>Takeaways/implications for practice</td>
<td><strong>It included samples of student work and verbatim teacher comments, which I find very interesting, informative and useful.</strong>&lt;br&gt;No examples to flesh these out or engage the reader and neither outlines the issues involved nor the strategies used, let alone any impact.</td>
</tr>
<tr>
<td>Easy to read/accessible</td>
<td><strong>The summary describes the key findings in a fairly accessible way that’s easy to read quickly</strong>&lt;br&gt;I think this kind of article is accessible to teachers</td>
</tr>
<tr>
<td>Structure</td>
<td><strong>The headings clearly signposted me to the information that I could read first (e.g. conclusions) and then decide whether to read more</strong>&lt;br&gt;Layout of text makes it possible for a teacher-reader to look at the first heading and the bullets under that heading and get the information they need about the research</td>
</tr>
<tr>
<td>Length/time</td>
<td><strong>I’m probably coming across as lazy here—but honestly, if I am looking to scan something over a cup of coffee, this is not going to hold my attention!</strong>&lt;br&gt;I would feel I’d need to set time aside in my day to digest this, which is a commitment I’d not often make without knowing what I’m really going to get out of it.&lt;br&gt;Teachers have little time (and inclination?) for reading.</td>
</tr>
<tr>
<td>Interesting/relevant topic</td>
<td><strong>I suspect that some subjectivity may have come in when reading the articles, as ones that interest me may have had my attention for longer, and therefore felt it conveyed it better</strong></td>
</tr>
<tr>
<td>Language</td>
<td><strong>The language of the summary is repetitive and difficult to follow</strong>&lt;br&gt;the writing is straightforward</td>
</tr>
<tr>
<td>Links/further information</td>
<td><strong>and there were many helpful links where you could and do further research if required</strong>&lt;br&gt;Teachers who might want to go further to ‘critique’ this paper could follow the reference to the project report (which is what I did—to see what evidence lay beneath this paper)</td>
</tr>
<tr>
<td>Context</td>
<td><strong>There wasn’t enough background context in this research</strong>&lt;br&gt;There is nothing to place it in context to indicate how it is located in what is known already and its value.</td>
</tr>
<tr>
<td>Trust/credibility</td>
<td><strong>I am assuming that a teacher’s purpose for reading this article (in this particular journal) is to learn from a ‘trusted researcher’ who (in his writing) shows expertise in both understanding students’ responses and in communicating important mathematical ideas.</strong></td>
</tr>
<tr>
<td>Methodology/use of evidence</td>
<td><strong>it was clear the research had been rigorous</strong>&lt;br&gt;But it is otherwise pretty useless as it doesn’t explain what a gain of, say, 30% might mean or give us much sense of how the research was conducted and only gives us the briefest hint of what the intervention involves.</td>
</tr>
<tr>
<td>Reflection/critical discourse</td>
<td>There are also suggestions for reflection or discussion.</td>
</tr>
<tr>
<td>Other/unsure</td>
<td><strong>It was interesting to compare different research and think about what makes it engaging to teachers in order to communicate findings successfully.</strong>&lt;br&gt;I really enjoyed this experience of using comparative judgement and would be interested in the outcomes of the research study!</td>
</tr>
</tbody>
</table>
likely that some items in the other/unsure categories may have overlapped, but were not counted as doing so, so this may be an underestimate. Cases where only a single item expressed a theme were also discounted in the overlap analysis.

Items that were outside of the overlap were as follows: mathematics teachers considered whether the topic was relevant or interesting to them to be more important than mathematics education researchers, who in turn considered the methodology or use of evidence, issues around trust and credibility, and the need for a research summary to prompt reflection or critical discourse more important than the mathematics teachers in the study. Although both groups considered language, teachers mentioned this theme more than researchers.

### DISCUSSION

#### Overall conclusions

Overall, both the group of mathematics teachers in the study and the group of mathematics education researchers in the study appear to agree amongst themselves to some extent when ranking education research summaries designed for teachers, measured by a Scale Separation Reliability (SSR) score of 0.61 and 0.68, respectively. When measured using comparative judgement, a moderate consensus was found between the two groups when ranking successful research summaries for teachers, with some important differences. The results of the questionnaire also support this view of moderate consensus with important differences.

#### Areas of consensus

If measured by the proportion of mentions when asked about features, mathematics teachers and mathematics education researchers agreed that graphic design was the most important element; that being summarised or focused, having takeaways or implications for practice and being easy to read or accessible and well-structured were key features of a
There is evidence here that mathematics teachers and mathematics education researchers agree on several elements when evaluating successful research summaries for teachers, suggesting the two communities may have some consensus in terms of what they consider useful in written research brokering products for teachers. The two groups in the study had considerable overlap in the themes they discussed when considering successful features of research summaries. This may suggest that (successful) research summaries, as part of the activity of knowledge brokering, could be considered boundary objects in the sense of bringing the activities of different communities into conversation with one another (Venkat, 2017). This may also reflect the study design, in which participants were asked to consider which of the research summaries communicated the research better to teachers—for the researchers, this may have meant placing themselves in the shoes of a teacher to the extent that they were able to identify the priorities and perspectives of a member of the teacher community successfully.

### Areas of difference

Mathematics education researchers mentioned methodology much more than mathematics teachers, which might be expected given that this is an important part of their professional role. They also referred to issues of trust and credibility more frequently, highlighting as we discussed in the literature review that knowledge brokering is dependent on the broker’s interpretative skill and the audience ‘blindly’ trusting this may be problematic (Shank & Brown, 2007), with some even discussing the contested nature of evidence (Aldridge et al., 2018). They also mentioned language and the importance of an interesting or relevant topic much less than mathematics teachers in the study, giving some suggestion of the community dissonance (Farley-Ripple et al., 2018) discussed earlier.
WHAT MAKES A GOOD EDUCATIONAL RESEARCH SUMMARY?

Conflicts and tensions

Beneath the headline areas of consensus and difference in this study, many interesting conflicts and tensions emerged. We have argued in this study that to help support partnerships between researchers and teachers, and to make this multidirectional knowledge flow effectively as well as ethically, research summaries should be as successful as possible at communicating the research to teachers, likely requiring specialist knowledge brokers to do so (Farley-Ripple & Grajeda, 2020). We have suggested this work is important but should be subject to evaluation, as research interpretation requires technical skill, communications expertise and ethical responsibility (H. Cooper & Hedges, 2009). We should also relate the characteristics of effective research summaries to their purposes, which may be varied, and the contexts in which they emerge.

The purpose of brokering educational research by creating research summaries is closely intertwined with the purpose of creating and using such research. Yet some of these purposes are in direct tension. For example, if the purpose of educational research summaries is to tell teachers how to be more effective, then they might have to be prescriptive summaries of findings of evaluation studies (ideally the ‘gold-standard’ RCT), focused on classroom practice. However, if the purpose of educational research summaries is to increase teacher autonomy, then they might have to include at least some detailed examples in context, contain principles rather than prescriptions, and allow space for teacher knowledge of their own pupils and context. Another example: if the purpose of the educational research summary is to be read and ideally acted on by as many teachers as possible—to prolifically stimulate

FIGURE 7 Proportion of items in each theme (relative frequency), teacher group against researcher group
immediate change in teachers’ actions—it should be summarised in ways that are short, accessible and well disseminated, with clear implications. But if the purpose of educational research summaries is to stimulate a more long-term change in teachers’ beliefs, it should be detailed, thought-provoking, prompt reflection, and allow space for deep consideration of the issues.

Similar arguments can be made for educational research summaries to be credible, useful, helpful in demonstrating to teachers their own paradigms, and indeed encourage teachers themselves to participate in or conduct research—and the list goes on, with many possible and sometimes conflicting purposes. No one research summary, or even type of research summary, can fulfil all these purposes. While some core features appear to exist which teachers and researchers agree on should be common to all, in other cases, features should be related to purposes in transparent ways. Features discussed in this study that were important to both groups and for which most of the response items were straightforward recommendations for good practice were:

- Good, clean, attractive graphic design including appropriate diagrams, ‘levels’ of text and effective use of space;
- Use of structure, for example boxes or bullet points, to highlight different sections and provide a visual ‘map’ of where to find different kinds of information;
- Focused summaries that are clear and accessible;
- Clear, concise and accessible language avoiding jargon and ‘waffle’.

**FIGURE 8** Significant (by frequency of mention) themes, overlapping and non-overlapping, for both groups
These findings are consistent with previous studies, such as Phipps et al. (2020).

Other themes were more ambiguous, reflecting more of a need for balance and purpose in research summary design, and tensions that might emerge from competing priorities in so doing. These themes included takeways or implications for practice, which participants generally agreed should usually be present but did not always agree on in terms of what they should be comprised of or how they should be framed.

We thus suggest that themes from the study that we have called ‘Goldilocks themes’ appeared to show more nuance and variability; we hypothesise that these reflect a type of underlying continuum in participants, likely reflecting tensions in purposes. These are shown in Table 8 with illustrations of the extreme ends of the continuum in each case and examples from the participants’ answers to the questionnaire to illustrate them.

These issues go deep, beyond the scope of this study: they reflect important tensions in the use of research in practice in education generally, questioning what constitutes useful and usable knowledge (Aldridge et al., 2018), who decides (Moss, 2017), and what kind of discourses are created around knowledge/s in education through the activities of knowledge brokering (Farley-Ripple & Grajeda, 2020). One participant suggested: ‘It is very difficult to be sure of …the purpose of communicating the research’, and we would tend to agree. We see clear links between the purposes and uses of educational research.
<table>
<thead>
<tr>
<th>Extreme A</th>
<th>Theme</th>
<th>Extreme B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Too summarised; not enough detail to judge rigour, applicability, or credibility e.g. ‘it was particularly reductive and one-dimensional and doesn’t really communicate the issues around using IWBs [interactive whiteboards]’.</td>
<td>summarised/focused (also related to length/time)</td>
<td>Not summarised enough; too much detail to absorb, not enough time or energy to read, not signposted enough e.g. ‘a lot of unnecessary information’.</td>
</tr>
<tr>
<td>No takeaways or implications for practice e.g. ‘Lots of information was of no real use to a classroom teacher. What would a teacher go and do differently after reading this?’</td>
<td>takeaways/implications for practice</td>
<td>Too focused on implications for practice or ‘neat’ takeaways without underlying theory, principles or mechanisms/models; too prescriptive; not nuanced enough e.g. ‘the “what works” culture is narrow-minded and fails to appreciate that experimental studies should be part of the research landscape, but not the end goal’.</td>
</tr>
<tr>
<td>No information on methodology or use of evidence e.g. ‘there is not enough information here to understand what was done and why’.</td>
<td>methodology/use of evidence</td>
<td>Too much confusing/distracting detail on methodology or analysis e.g. ‘The presentation of results of the studies is tedious and unnecessary’</td>
</tr>
<tr>
<td>No branding, organisational provenance or authorship details available; no consistency between research summaries from the same source e.g. ‘Authors of the other excerpt are unknown to me, thus I’d spend time trying to find out more about them’</td>
<td>trust/credibility</td>
<td>Misuse/overextension of credibility; commercial/political interests; blind trust in expertise without critical feedback e.g. ‘I wondered to what extent trust in the author of an article has an effect’</td>
</tr>
<tr>
<td>No links or further information provided links do not work or are outdated e.g. ‘there were many helpful links where you could and do further research if required’.</td>
<td>having links/further information available (also related to trust/credibility and methodology/use of evidence)</td>
<td>Links and further information overwhelming and off-putting; too much detail e.g. ‘there is no guarantee the links will be worth the effort’</td>
</tr>
<tr>
<td>No detail on context or mechanisms/models in relation to context; e.g. ‘There is nothing to place it in context’</td>
<td>addressing context</td>
<td>Discussion of specific contexts but no examination of generalisability/wider impact and the mechanisms and models which might underlie this e.g. ‘The sample was taken from only three schools in the same area and “nearly all of them” were from the same disadvantaged background how does this actually compare to nondisadvantaged students?’</td>
</tr>
<tr>
<td>Too short to be useful; ‘dumbed down’ e.g. ‘it is just too brief’</td>
<td>length/time</td>
<td>Too long, detailed, unstructured, hard to read, or impenetrable to be useful e.g. ‘As a teacher, I find it hard to read long research papers due to time constraints’</td>
</tr>
</tbody>
</table>
summaries and the need to resolve some of these tensions in design, which we summarise in Table 9.

**IMPLICATIONS**

**Implications for knowledge brokers**

In the literature review section, we found that community dissonance between teachers and researchers may be reduced by promoting better relationships and partnerships between them, and ensuring the knowledge flow is not linear (e.g., H. Davies et al., 2008). We argued that research summaries for teachers should be a key part of this process, serving as boundary objects in a non-linear, multidirectional model of knowledge flow, and as part of a wider partnership model that helps to break down hierarchies between researchers and practitioners in education. One implication of the results of the follow-up questionnaire is that teachers might benefit from more development and support in areas researchers found important: the methodology or use of evidence, issues around trust and credibility, and the need for a research summary to prompt reflection or critical discourse. In turn, researchers may benefit from development and support in areas teachers considered more important: choice of interesting/relevant topic, structure and language. Since knowledge brokers in education work to help teachers and researchers understand each other's worldviews emanating from the different types of work they do (A. Cooper et al., 2020), evidence from studies such as this one may be valuable in underpinning such activity.

But, as we have discussed previously, there are power structures in educational systems at play that mean these relationships are not always balanced (e.g., Lysenko et al., 2014). Research summaries should serve as boundary objects, allowing teachers and researchers—and other stakeholders in the knowledge brokering process—to communicate better with one another, and crucially allowing teachers more agency and space to respond critically to ideas around educational evidence. One encouraging result of this study is that researchers found ideas of reflectiveness and prompting critical discourse to be important when considering features and factors of research summaries. For example, one said that the way the research summary was constructed ‘allows the teacher to be more constructively critical of it’; another that ‘It invited teachers to consider students’ actions and talk as sources of evidence for reflection’. Whether this happens in practice, and is built into the design cycle of such knowledge brokering products, is an important and unresolved question.

The findings from this study imply that the characteristics of effective research summaries should be explicitly related to their purposes, which may be varied, and that the work of knowledge brokers producing such research summaries should be subject to greater evaluation using critical frameworks, which take account of (at least) the perspectives of both teachers and researchers.

One recommendation that arises from this work suggests educational research summaries should have clearly stated purposes, alongside a statement of conflict of interest, and space for dialogue, critique and feedback from teachers as part of a non-linear model of knowledge flow. However, one criticism of this approach is that, rather than considering purposes of research use, in line with a more multi-perspective view, *practices* should be considered instead because of ‘a strong need to conceptualize practices associated with use due to the complex nature of decision-making in education in which technical, political, and instructional challenges abound’ (Farley-Ripple et al., 2018, p. 9). Further work in this area is suggested to resolve some of these contentions, and combine the ideas of ‘Goldilocks’ features with purposes to create a detailed framework for evaluation which can then be tested.
### Table 9: Purposes and hypothesised characteristics of research summaries for teachers

<table>
<thead>
<tr>
<th>Theme from study</th>
<th>If educational research summaries are...</th>
<th>Then they should be...</th>
<th>And they should not be...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summarised/ focused (also related to length/time)</td>
<td>For increased awareness (Hutchinson &amp; Huberman, 1994) an introduction to research (Phipps et al., 2020) to provide insights into students' conceptions and misconceptions and to describe learning trajectories for some mathematical topics (Lobato &amp; Walters, 2017)</td>
<td>Easy to read, accessible ‘research story’ with clear explanation of new vocabulary, placed in context</td>
<td>Too complex or detailed; lack of definition of terms; too much content that is contested/unresolved</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Subject-specific, focused on mapping and communicating learning trajectories and topic-specific; detailed</td>
<td>Vague, missing clear learning trajectories, missing examples of what teachers or students might say or do; not enough detail; too summarised or compressed</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Takeaways/ implications for practice</td>
<td>For instrumental use: as learning how to apply modified or new practices (Hutchinson &amp; Huberman, 1994) Professional knowledge for teaching: knowledge about how students in a given class are likely to respond to particular aspects of instructional tasks plus knowledge of patterns that are observed across classes... building a repertoire of possible student responses to particular instructional tasks (Cai et al., 2020) To give teachers the ability to make informed choices about alternatives to current practice (Hutchinson &amp; Huberman, 1994) For conceptual use as learning or enlightenment; as the acquisition of new perspectives or attitudes (Hutchinson &amp; Huberman, 1994)</td>
<td>Clear description of modification or changed practice, with evidence of impact and contextual information; design principles for new practice Analyses of a range of examples of student response Clear comparison of ideas and implications, with balanced advantages and disadvantages in context Clear explanations of concepts, theories or ideas and how they may fit together to form a possible narrative, along with a wider ‘lens’ or paradigm explanation</td>
<td>Lack of contextual information; too simplistic in recommending new practice Lacking in examples or useful patterns of responses (i.e. applied theory) Commercially/politically motivated or strongly biased; neglecting the ‘why’ as well as the ‘what’; disempowering</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Prescriptive, too focused on implications, lacking in principles or paradigms; lacking in clear explanation of what it is intended to replace and why it is better.</td>
</tr>
<tr>
<td>Addressing context</td>
<td>A framework for teachers to make sense of data they themselves collect from observing their own students that can be used as a rationale for making further instructional decisions (Cai et al., 2020)</td>
<td>Clear representations of relevant and recent theory; detailed conceptual maps with models attending to contextual factors</td>
<td>Prescriptive; politically motivated; disempowering; failing to address contextual factors</td>
</tr>
</tbody>
</table>
Another recommendation is that graphic design is worthy of ‘serious and critical attention’ (Barnard, 2013, p. 8). To date, not enough attention has been paid to design elements in research summaries, and in particular how they interact with structural elements to signpost levels of importance, break text and images into sections to support storytelling, and increase accessibility. There is evidence to suggest that use of accessible language is important when designing research summaries, and this study adds to the weight of such evidence. We found that teachers considered clear language, structure and ease of reading to be important, and this aligns with evidence considered in the literature, which suggests that the language used in research papers is inaccessible and requires brokering (See et al., 2016); however, consistent with the idea of a community dissonance, mathematics education researchers considered language less important than mathematics teachers. The question of who is doing the work of knowledge brokering and whether they are able to understand and act on the perspectives of both teachers and researchers remains therefore a key focus of future study. The work of knowledge brokers is important and needed, but may be susceptible to commercial and political bias, and, because trust is implicit in the knowledge brokering processes of selection, interpretation and faithful dissemination of evidence, robust frameworks for evaluation of research summaries should be developed, building on ideas in this study, which allow for the perspectives and priorities of teachers and researchers.

Implications for use of CJ

The broad question of what makes a piece of good ‘work’ seems to align well with CJ as a methodology; on the other hand, the question will always be ‘good at what?’ and ‘for whom?’ For example, there could be an argument presented that previous CJ research on marking student work has offered up high rates of reliability because of decades of teacher compliance; the practice of being trained to read marking criteria so closely constraining ideas of ‘quality’ such that teachers feel cannot, for all sort of reasons, deviate from it even when given an opportunity. Similarly, participants in this research may have used proxies for quality in research summaries—such as successful graphic design or length—to make decisions, based on orthodoxies in the discourse around educational evidence or unexamined heuristics. Thus, these results should be interpreted as exploratory, and further work in this area is needed.
This work also furthers the development of the idea of flexible assessment criteria supporting diversity in assessment (e.g., Seery et al., 2012), as it used CJ to assess a different type of artefact than had been used previously; criticisms of this may suggest that the artefacts used here were not sufficiently similar or designed to fulfil similar enough purposes to retain validity. Further studies along similar lines are likely to shed more light on this issue.

It is likely that some participants in this study considered themselves knowledge brokers and more research on this role and its challenges/limitations would be valuable. An exploration also of whether participants enjoyed the process of using CJ or found it to be preferable to using received evaluation criteria might be of interest to the field. Although we did not ask them this explicitly, many participants in both groups in this study reported enjoyment and interest from engaging with the CJ task, but other studies, such as Verhavert et al. (2018), reported that participants found it monotonous.

Comparative judgement has the potential to have further impact in the field of knowledge brokering evaluation, and in particular comparing the results of this study with a larger one involving a group of knowledge brokers themselves (although defining this group is not easy) has the potential to illuminate this area further. This study suggests that there is potential to use CJ as a basis for teacher and/or researcher professional development in a similar way to McMahon and Jones (2015), who suggest the activity of peer assessment via CJ is in itself a valuable teaching and learning activity for pupils. In particular, asking participants both within groups and across groups to complete the CJ task together in pairs or groups may be fruitful in exploring further some of the issues mentioned here.

CONFLICT OF INTEREST
The first author wrote some of the materials (Cambridge Mathematics Espressos) evaluated in the study. The second author has no conflict of interest to declare.

ETHICAL APPROVAL
The study was reviewed and received approval by the Research Ethics Committee of the University of Cambridge Faculty of Education.

DATA AVAILABILITY STATEMENT
Under the terms of consent provided in the study, participants did not give permission to make the data publicly available.

ORCID
Lucy Rycroft-Smith 🐳 https://orcid.org/0000-0002-6058-0946
Andreas J. Stylianides 🐳 https://orcid.org/0000-0002-3526-0342

ENDNOTE
1 https://cjscaling.cambridgeassessment.org.uk/

REFERENCES


---