Educ Stud Math
DOI 10.1007/s10649-017-9750-y

Teachers' appraisals of adjectives relating to mathematics tasks

Colin Foster[1] · **Matthew Inglis**[2]

Abstract Curricular implementations are unlikely to deliver the anticipated benefits for mathematics learners if written guidance to teachers is interpreted and enacted differently from the ways that policymakers and curriculum designers intend. One way in which this could happen is in relation to the mathematics tasks that teachers deploy in the classroom. Teachers and curriculum designers have developed an extensive vocabulary for describing tasks, using adjectives such as 'rich', 'open', 'real-life', 'engaging' and so on. But do teachers have a shared understanding of what these adjectives mean when they are applied to mathematics tasks? In study 1, we investigated teachers' appraisals of adjectives used to describe mathematics tasks, finding that task appraisals vary on seven dimensions, which we termed engagement, demand, routineness, strangeness, inquiry, context and interactivity. In study 2, focusing on the five most prominent dimensions, we investigated whether teachers have a shared understanding of the meaning of adjectives when applied to mathematics tasks. We found that there was some agreement about inquiry and context, some disagreement about routineness and clear disagreement about engagement and demand. We conclude that at least some adjectives commonly used to describe tasks are interpreted very differently by different teachers. Implications for how tasks might be discussed meaningfully by teachers, teacher educators and curriculum designers are highlighted.

Keywords Mathematical knowledge for teaching · Task appraisals · Task characteristics · Task design in mathematics

✉ Colin Foster
colin.foster@nottingham.ac.uk

[1] Centre for Research in Mathematics Education, School of Education, University of Nottingham, Nottingham NG8 1BB, UK

[2] Mathematics Education Centre, Loughborough University, Loughborough, UK

Published online: 09 March 2017

1 Introduction

Why do mathematics curricular implementations so often fail to deliver the anticipated benefits for learners (Cuban, 1993)? One reason may be that written guidance to teachers is interpreted and enacted in different ways from the intentions of policymakers and curriculum designers (Remillard & Bryans, 2004; Remillard & Heck, 2014). For example, mathematics tasks for the classroom are frequently described, both informally and in academic papers and policy documents, using a plethora of adjectives, such as 'authentic', 'rich' and 'complex' (Shimizu, Kaur, Huang, & Clarke, 2010, p. 3). The US Common Core State Standards for Mathematics, for instance, advises that mathematical understanding and procedural skill can be assessed using mathematics tasks "of sufficient *richness*" (CCSSI, 2010, p. 4, emphasis added). Similarly, the English National Curriculum suggests offering learners "rich and sophisticated problems" (DfE, 2014, p. 3). But how are terms such as 'rich' tasks interpreted by teachers? Do teachers have a shared understanding of what these adjectives mean when they are applied to mathematics tasks? Academics, policymakers and curriculum designers who use such language to communicate their intentions will succeed only if teachers interpret these terms in the ways intended.

The mathematics tasks offered to students are fundamental to their learning (Watson & Ohtani, 2015); indeed, Sullivan, Clarke, and Clarke (2013, p. 57) argued that "the nature of teaching and what students learn are defined largely by the tasks that form the basis of their actions". Thus, helping teachers to design and select high-quality mathematics tasks for use in the classroom is critical for improving students' learning of mathematics (Jones & Pepin, 2016). However, little is known about teachers' interpretations of adjectives describing mathematics tasks and, indeed, about teachers' language about tasks more generally.

Researchers have developed language to categorise and describe desirable features of mathematics tasks, typically by defining words such as 'rich' (e.g., Ahmed, 1987; Griffin, 2009; McDonald & Watson, 2010; Stylianides & Stylianides, 2008; Swan, 2008; Yeo, 2007, 2017). Whilst different definitions of a word like this usually have many commonalities, they are sometimes contrasting, or at least non-identical.[1] Curriculum documents, such as those quoted above, and teacher resources tend to use some of the same adjectives, but without explicit definitions, so that it is implicitly assumed that there is a language shared by the teachers who are the intended readers of these documents. If this is not the case, the intended and enacted curricula may not be the same, threatening fidelity of implementation.

In this paper, we explore how teachers appraise adjectives relating to mathematics tasks. Research into teacher beliefs and perspectives (see, for example, Bräunling & Eichler, 2015) gives us reason to suppose that teachers might have different understandings of adjectives relating to tasks, although little is known specifically about how teachers appraise adjectives relating to mathematics tasks (Clarke, 2013; Clarke & Mesiti, 2010; Lengnink & Prediger, 2003; Philipp, 2007; Simon & Tzur, 1999). So our research strategy is to explore two possible ways in which teachers may disagree about the meaning of adjectives used to appraise mathematics tasks. Disagreement could be about:

(a) How different adjectives relate to each other (e.g. are tasks perceived to be 'enjoyable' also perceived to be 'engaging'?) or
(b) How particular adjectives latch on to specific tasks

[1] For example, Ahmed (1987) stressed that rich activity involves accessibility, extendibility, the potential for surprise, enjoyment and originality, whereas Swan (2008) emphasised multiple solution pathways and student collaboration in which students shift roles and explain and teach one another.

Of course, these types of disagreement are not entirely independent. For instance, if there were agreement in the sense of (b), it is hard to see how there could not also be agreement in the sense of (a). But it seems possible for there to be agreement in the sense of (a) but not (b); for example, perhaps two teachers both always think that 'engaging' tasks are also 'enjoyable', but routinely disagree about whether or not specific tasks are 'engaging'.

Our aim in the study reported here was to characterise what we might call the 'task appraisal space', consisting of the space built by the dimensions in which tasks can be appraised. We ask: Is it sensible to describe the task appraisal space in terms of a dimensional structure? If so, what is that structure (i.e., what are its dimensions)? Thus, first, we investigated whether adjectives for mathematics tasks tended to cluster together and, if so, in what ways. In doing so, we estimated the number of dimensions upon which mathematics tasks can be appraised. Second, we investigated whether teachers have a shared understanding of the meaning of adjectives that are used to describe tasks. Given a specific task, do teachers agree about whether it is, for example, 'rich'?

2 Appraisals of mathematics tasks

Christiansen and Walther (1986) regarded a classroom mathematics *task* as the prompt given to the students, whereas the *activity* comprised the particular students' responses to that prompt. However, it is widely acknowledged that it can be very difficult to separate the task from its resulting activity (Watson & Mason, 2007). Thus, although many adjectives can be used to describe mathematics tasks, it can often be difficult to say to what extent the quality lies inherently within the task as given (i.e., the prompt) and to what extent it is a property of the way particular students react in a particular instantiation of the task on a particular occasion. The distinction between a *property* of a task and an *effect* of a task is not always clear-cut. An 'engaging' task might be regarded as such in isolation, or, alternatively, as not truly engaging until it engages some particular student(s) in a classroom somewhere (Mason, 2016, p. 276). So when a teacher judges a task to be 'engaging', it may be through recalling the task enacted in a classroom at some point in the past or else imagining some typical classroom and the anticipated response of a typical hypothetical student. According to Mason (2015, p. 15), "There are no rich mathematical tasks, only tasks used richly"; nevertheless, it is common practice, as shown in the curriculum documents quoted above, to regard qualities such as 'richness' as properties of the task rather than only the activity. Given these difficulties, we used 'task/activity' to refer to a specific task which has been enacted in the classroom and 'task' to refer to a written prompt which we present to teachers without assuming that they will have witnessed an enactment of it.

Our approach to understanding teachers' appraisals of adjectives relating to mathematics tasks can be situated within the *empirical semantics* tradition developed by the Oslo Group of philosophers in the early to mid-twentieth century (e.g., Gullvåg, 1955; Næss, 1938). Næss and his followers suggested that a word's meaning can be inferred only from studying the way it is used. Gullvåg (1955), for instance, pointed out that any suggestion of what a word means:

> is merely an unsupported guess as long as no systematic testing of it has been attempted. In order to test it systematically it is hardly sufficient that a single person registers his own reactions to this or that sentence, or makes pronouncements based on intuitions, or undertakes scattered observations of others' usage. (p. 343)

Instead, systematic observations are necessary. In drawing on this tradition of empirical semantics, it is not necessary that participants use words spontaneously, only that understanding of the meaning of words derives from their use. For example, Næss (1938) used questionnaires to ask for examples of things that are 'true' and from this determined use of the word 'true'.

Teachers in the same educational system work in schools with related curricula and exchange media which use similar words to describe tasks. In this way, ideas are transferred from research and design to schools, and teachers meet and talk to one another, face to face and online. Even if individual interpretations of tasks are different, teachers understand each other through the language about tasks that they use. Hence, it may be that the way teachers talk about tasks can be characterised as being built on a common space in which they appraise tasks.

3 Study 1: Interpretations of task appraisals

In study 1, we empirically investigated the task appraisal space by asking: Is it sensible to describe this space in terms of a dimensional structure? If so, on how many broad dimensions are mathematics tasks/activities perceived to vary? We chose to answer this question by using an exploratory factor analysis (EFA), a statistical approach that seeks to represent a large number of variables using a smaller set of factors whilst accounting for as much of the original variance as possible. The technique works by looking at the matrix of correlations between the original variables. If a subset of these variables are all strongly intercorrelated, then they can be thought of as representing a single underlying construct. These constructs would reveal the ways in which adjectives cluster together in teachers' appraisals. We did not assume beforehand that it would be possible to describe the space in terms of any particular underlying dimensional structure, but rather sought to determine empirically whether this was the case.

3.1 Method

The study was approved by Loughborough University's Ethics Approvals (Human Participants) subcommittee.

3.1.1 Stimuli

We created a list of 84 adjectives, given in Table 1, which have been used to describe mathematics tasks. We obtained these adjectives by searching teacher professional journals and academic journals for articles relating to mathematics tasks and task design and collating the adjectives found. In addition, we reviewed two online discussion boards used by mathematics teachers and several teacher-facing websites for adjectives relating to mathematics tasks, and we also sought suggestions from local teachers and mathematics education academics on any words that they felt should be included. An inclusive approach was taken, so that when in doubt about whether to include a particular word or not, it was generally decided to include it. We assumed that, in this way, we would cover a broad variety of possible adjectives, so that most of the words that teachers would wish to use would be present in our selection. Because the adjectives had been obtained from places such as mathematics teacher professional journals and forums, it is plausible that teachers sometimes need to interpret them in order to make sense of educational discussions. In this way, we intended the instrument to have ecological validity.

Table 1 The 84 adjectives used in study 1

absorbing	deep	intriguing	real-life
abstract	demanding	investigative	realistic
accessible	difficult	irrelevant	repetitive
active	distracting	meaningful	rich
amusing	dull	mechanical	routine
analytical	easy	memorable	rule-based
appealing	engaging	modelling	serious
applied	enjoyable	motivating	shallow
attention-grabbing	exciting	multistep	simple
authentic	experimental	open	slow
boring	exploratory	original	stimulating
challenging	extendable	passive	strange
closed	formal	perplexing	surprising
collaborative	fun	pleasing	technical
complicated	hands-on	practical	tedious
conceptual	ICT-based	predictable	testing
concrete	imaginative	problem-solving	thought-provoking
confusing	informal	problematic	time-consuming
context-based	inquiry-based	procedural	unusual
cooperative	inspiring	puzzling	useful
creative	interesting	quick	useless

3.1.2 Participants

We asked teachers to provide spontaneous ratings of tasks in a somewhat artificial setting—that of an Internet research study—in order to allow us to canvass a large number of teachers. Participants were teaching professionals recruited by sending speculative e-mails to teachers known to several UK higher education institutions. We also posted a notice about our survey on several online discussion forums frequented by teachers and asked various individuals to mention the study on their blogs and Twitter accounts. A UK mathematics teacher professional journal mentioned the study in their editorial and another UK mathematics teacher professional association mentioned the study in its e-mail news bulletin, as did a large UK mathematics curriculum development organisation. In each case, snowball sampling was employed, with recipients being asked to forward details of the study to any other mathematics teachers with whom they were in contact. The recruitment messages briefly explained the purpose of the task and invited participants to visit a given web address if they wished to take part. A total of 360 participants completed the study.

We recorded the number of years of teaching experience for each participant, together with the age range they normally taught. A large majority, 92%, of our participants were secondary school teachers (teaching children aged 11–18 years), and most had between 5 and 25 years teaching experience ($M = 14.8$, SD = 10.1).

3.1.3 Instrument

Having received the invitation, participants clicked through to a website where the purpose of the study was again explained to them and their demographic information was recorded. Participants were then given the following prompt:

> Please think of a *particular* mathematical task/activity that you used recently with students, or you saw another teacher use with students. Keeping this specific task/activity in mind, please use the rating scale below to describe how accurately each word

in the table below describes the task/activity. Describe the task/activity as it was used, not how it could have been used if improved or adapted. (original emphasis)

They were then presented with a list of the 84 adjectives in Table 1, given in a random order. Participants were invited to respond on a five-point Likert-type scale (very inaccurate, moderately inaccurate, neither inaccurate nor accurate, moderately accurate and very accurate). Finally, participants were thanked for their time and invited to contact the research team if they wanted further information.

It is common practice within empirical semantics to elicit the meaning of words by inviting participants to make appraisals. As mentioned above, Næss (1938) asked participants for examples of things that are true in order to determine their use of the word 'true', and others within the empirical semantics tradition have asked whether situations or objects are well described by words. For example, Knobe and Fraser (2008) investigated causation by offering participants fictional scenarios and asking whether they believed that they were well described by causal statements, and Pettit and Knobe (2009) did something similar for intentional action.

It is clear that the approach taken here does not permit us to uncover the total richness of the thinking of a single teacher since they could not choose their own adjectives but were restricted in their assessments to the 84 adjectives we chose. However, by asking 360 teachers of varied levels of experience to think about a recent task that they had used or seen used, we intended to expose the collective richness of teachers' thinking, allowing them to draw on a vivid recollection of a task of their choosing. The details of the specific tasks were unimportant for the purposes of the analysis, provided that each teacher brought to mind one specific task that he/she could recollect in sufficient detail. We relied on the assumption that teachers would think of a broad variety of possible tasks so that we would expect to capture most kinds of appraisal and obtain the variation that we needed for the EFA to be successful. By allowing teachers to choose a task that they had used or seen recently, we expected that they would be more likely to be able to rate it accurately on the five-point scale, allowing us to obtain reliable data.

3.2 Results

The suitability of the data for factor analysis was first evaluated. The Kaiser-Meyer-Olkin value was .902, falling within the 'marvellous' range according to Kaiser (Dziuban & Shirkey, 1974). Bartlett's test of sphericity [$\chi^2(3486) = 16{,}917$, $p < .001$] confirmed that the correlation matrix contained non-zero terms. Together, these constitute evidence for the presence of adequate structure within the task appraisal space for performing an EFA. Hence, participants' ratings were entered into an EFA using the maximum likelihood method and an oblimin rotation. Horn's parallel analysis (Horn, 1965) suggested that seven factors should be extracted, accounting for 21, 6, 5, 5, 4, 2 and 2% of the variance, cumulatively accounting for 44% of the variance. The factor loadings are presented in the Appendix, where a high positive loading indicates that the adjective is very representative of that factor, a high negative loading indicates that it is very unrepresentative of that factor, and a zero loading indicates that it is independent of that factor. The highest-loading adjectives in each factor are shown in Table 2.

Following inspection of the factor structure, we named the seven factors as follows: engagement, demand, routineness, strangeness, inquiry, context and interactivity. We tried to choose names that encompassed the sense of as many of the highest-loading adjectives as

Table 2 Adjectives loading with magnitude >.5 for each factor

Engagement	Demand	Routineness	Strangeness	Inquiry	Context	Interactivity
enjoyable	difficult	routine	strange	open	real-life	hands-on
fun	complicated	repetitive		inquiry-based	realistic	cooperative
pleasing	demanding	procedural		deep	context-based	collaborative
appealing	perplexing	formal		exploratory	applied	practical
attention-grabbing	easy[a]	mechanical		investigative		
motivating	challenging	rule-based		rich		
stimulating	simple[a]			thought-provoking		
memorable	problematic			closed[a]		
boring[a]	puzzling			analytical		
interesting						
absorbing						
exciting						
inspiring						
dull[a]						
engaging						

[a] Indicates negatively loaded

possible. In all cases except the first and last factors, we used a name based on one of the top three adjectives within that factor. The relationships between the different factors are shown in Table 3, and, as expected, these correlations were all fairly low.

3.3 Discussion

We now discuss each factor in turn.

3.3.1 Engagement

The adjectives loading strongly onto this factor seemed not to be explicitly mathematical, and they primarily related to positive affect. We termed this factor 'engagement' on the grounds that it appeared to relate to students' participation and willing involvement in the task/activity.

3.3.2 Demand

This factor related to mathematical demand, with high loadings for 'difficult', 'complicated', 'demanding' and 'perplexing'. The adjectives 'easy' and 'simple' loaded negatively. Table 3 shows that the demand factor was unrelated to the engagement factor ($r = .08$), indicating that

Table 3 Correlations between each pair of factors

	Engagement	Demand	Routineness	Strangeness	Inquiry	Context	Interactivity
Engagement	1.00	.08	−.20	−.04	.32	.26	.29
Demand		1.00	.02	−.06	.28	.03	.01
Routineness			1.00	.06	−.10	.05	−.13
Strangeness				1.00	.09	−.06	−.13
Inquiry					1.00	.30	.20
Context						1.00	.27

how engaging or enjoyable a task/activity was perceived to be was largely independent of its perceived level of challenge.

3.3.3 Routineness

The adjectives 'routine', 'repetitive', 'procedural' and 'mechanical' loaded cleanly onto this factor. Routineness seemed to be only weakly related to engagement ($r = -.20$), meaning that although routine tasks/activities were less likely to be engaging than non-routine tasks/activities, this relationship was weak and we would expect to be able to find engaging routine tasks/activities.

3.3.4 Strangeness

Few adjectives loaded cleanly onto this factor (only 'strange' had a loading over .5). The factor appeared to relate to unusual aspects of tasks/activities. Of the six adjectives which loaded most strongly onto this factor, four ('useless', 'irrelevant', 'distracting' and 'strange') were amongst the 15 adjectives with the lowest mean ratings. In other words, few teachers felt that the tasks/activities that they were thinking of were well described by the adjectives most associated with this factor. Although teachers do not seem to think of strange tasks when asked, that does not mean that these tasks do not exist. Therefore, theoretically, it makes sense to retain this characteristic in the space, although since there is only one item loading more than .5 on this factor, it would be difficult to assess where specific tasks would fall on it.

3.3.5 Inquiry

This factor was characterised by adjectives such as 'open', 'inquiry-based', 'exploratory', 'investigative' and 'rich'. Although tasks/activities which are high on the inquiry factor seem more likely to be engaging, this relationship was weak ($r = .32$), suggesting that there is no automatic link between the use of inquiry-based tasks/activities and student engagement.

3.3.6 Context

The adjectives which loaded strongly onto this factor related to the real world ('real-life', 'realistic', 'context-based' and 'applied'). Tasks/activities which rated highly on the context factor were slightly more likely to rate highly on the engagement factor ($r = .26$) and the inquiry factor ($r = .30$), but again, these relationships were weak.

3.3.7 Interactivity

The final factor seemed to encompass ways in which students interact both with one another ('cooperative' and 'collaborative') and with practical equipment ('hands-on' and 'practical'). We felt that this combination was best described by the name 'interactivity' rather than by any of the individual adjectives that loaded onto the factor. Tasks/activities which were rated highly on the 'interactive' factor were more likely also to be rated highly on the engagement ($r = .29$), inquiry ($r = .20$) and context ($r = .27$) factors, although all of these relationships were weak.

3.4 Summary

Our EFA revealed that teachers' task appraisals varied on seven factors: engagement, demand, routineness, strangeness, inquiry, context and interactivity. Roughly speaking, then, the teachers in our study characterised tasks by positioning them within a seven-dimensional space.

4 Study 2: Between-teacher consistency

The results from study 1 showed a degree of within-teacher consistency in how adjectives are judged to characterise mathematics tasks. Indeed, it was possible to describe the task appraisal space with a dimensional structure only because there was sufficient within-teacher consistency to make this statistically feasible. By 'within-teacher consistency', we mean the extent to which teachers give similar responses to adjectives from the same factor, so, for example, if there were high within-teacher consistency, then we would expect that teachers who rate a task as being enjoyable would also rate it as being fun, and teachers who rate a task as being not enjoyable would also rate it as not fun. However, these data cannot address an important related question, that of *between-teacher* consistency. By this, we mean the extent to which different teachers give similar responses to the same task; that is, if one teacher believes that a task is 'rich', will another teacher agree? In study 2, we investigated these issues by examining whether teachers rate specific tasks in a homogeneous or heterogeneous fashion.

4.1 Method

In order to explore the extent to which the appraisals of mathematics tasks are shared amongst teachers, in study 2 we adopted a two-stage approach. In the first stage, we asked teachers to consider one specific task and to rate how well described it would be by a series of adjectives, chosen to best represent five of the factors found in study 1 (we explain below the reason for using only five of the factors). In the second stage, we adopted a slightly different approach and asked teachers to look at two tasks and decide which could be better described by each adjective. This two-stage approach allowed us to assess whether teachers' task appraisals were consistent in both absolute and relative terms.

We created a short 'mathematics task scale', which we used to quantify teachers' appraisals of the task. We excluded the strangeness factor, since it did not have enough adjectives with sufficiently high loadings, and also the interactivity factor, as we felt that the adjectives associated with that factor seemed to describe activities rather than tasks. From each of the remaining five factors, we used the four adjectives with the highest loadings, except that we omitted 'procedural' from the routineness factor (using 'mechanical' instead—the fifth highest loading adjective) as 'procedural' also loaded reasonably well onto the inquiry factor. This gave 20 adjectives in total.

The study was approved by Loughborough University's Ethics Approvals (Human Participants) subcommittee. We again adopted an Internet-based method. Participants were recruited in a similar manner to study 1 and, after giving their demographic information, were presented with the 'Trapped Squares' task shown in Fig. 1. To situate the task, participants were told that it was "aimed at a typical Year 9–10 class (ages 13–15)." Participants were then asked to "Select how accurately each of the following words describes *this task*" (original emphasis).

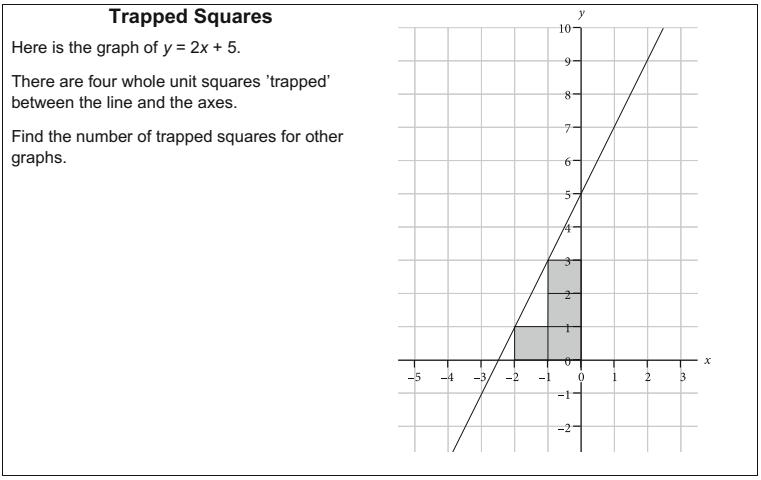

Fig. 1 'Trapped Squares' task (Foster, 2011)

The 20 adjectives given in Table 4 were presented in a random order and participants were invited to respond on the same five-point Likert-type scale as in study 1. Once participants had completed the first stage of the study, they were presented with the 'Factors' and 'Product' tasks shown in Fig. 2. They were asked to "say which task is better described by each of these adjectives", again being presented with the 20 adjectives from Table 4 in a random order. Responses were recorded on a five-point Likert-type scale (definitely the 'Factors' task, probably the 'Factors' task, the adjective describes both tasks equally well, probably the 'Product' task and definitely the 'Product' task). For example, participants were asked to say which task (the 'Factors' task or the 'Product' task) is better described by the adjective 'routine', giving one response from the five options listed above. Once again, participants were told that both tasks were aimed at a typical class of 13- to 15-year-olds.

We did not regard the specific details of the tasks chosen for this study as crucial since here we were concerned only with teachers' consistency. The three tasks used in the study were chosen from a series of tasks designed for secondary school students (Foster, 2008, 2010, 2011, 2013). From this large bank of 240 tasks, we eliminated all those which could not be presented in a single webpage and which were not aimed at 13- to 15-year-olds. The desire for tasks that would be quick for participants to comprehend limited the number of contextual tasks considered—an important limitation. From the remaining 50 tasks, we randomly selected the three shown in Figs. 1 and 2.

Clearly, the teachers' attribution of properties might be expected to depend heavily on the classroom context in which they imagined using the task; for example, the demand of a task will vary depending on the previous experiences of the student, as will other properties,

Table 4 The 20 adjectives used in study 2, organised by factor

Engagement	Demand	Routineness	Inquiry	Context
enjoyable	difficult	routine	open	real-life
fun	complicated	repetitive	inquiry-based	realistic
pleasing	demanding	formal	deep	context-based
appealing	perplexing	mechanical	exploratory	applied

Product	Factors
1 2 3 4 5 6 7 8 9 Use the digits 1 to 9 once each to make two numbers which, when multiplied together, give the largest possible product.	Find some numbers with exactly 4 factors. Find some numbers with exactly 5 factors. Find some numbers with exactly 6 factors. What can you find out about numbers that have different numbers of factors?

Fig. 2 'Product' task and 'Factors' task (Foster, 2013)

depending on the kind of pedagogy envisaged. However, our sampling method makes it unlikely that there would be systematic bias amongst the teachers within our sample towards particular teaching styles or classroom characteristics.

A total of 102 participants completed the study. As in study 1, a large majority (all except one) of our participants were secondary school teachers (teaching children aged 11–18 years), and most had between 5 and 25 years teaching experience ($M = 15.0$, $SD = 21.7$). Nine participants who had more than one missing response were excluded, and another nine participants who were missing a single response received imputed values (equal to the mean of the responses for that item from the rest of the data). This left a total of 93 participants whose data were analysed.

4.2 Results

We first assessed the internal reliability of each of the five factors by calculating Cronbach's alphas, which measure internal consistency by indicating how closely related a set of items are as a group. In our case, high alphas would imply that teachers gave similar responses to each of the adjectives that made up a single factor. The alphas for each factor in both stages are shown in Table 5. The alphas for engagement and inquiry comfortably exceeded Cohen and Swerdlik's (2009) minimum recommendation of .65, and the alpha for demand was acceptable, but the alphas for routineness and context fell slightly short of this recommended value. Consequently, our findings with respect to routineness and context should be regarded more cautiously than those relating to the other factors.

4.2.1 Stage 1 results

We summed each participant's responses to each item on each factor and subtracted 4, which resulted in a score from 0 to 16 on each factor for each participant. Thus, a participant who scored 0 on the engagement factor would have felt that the task was highly non-engaging, whereas a participant who scored 16 on the same factor would have felt the reverse.

Table 5 Cronbach's alpha values for the scales used in the two stages of study 2

Factor	Cronbach's alpha	
	Stage 1 ($N = 93$)	Stage 2 ($N = 93$)
Engagement	.947	.908
Demand	.663	.828
Routineness	.610	.585
Inquiry	.728	.806
Context	.602	.641

Histograms for each factor, along with the mean and standard deviation of participants' scores, are shown in Fig. 3.

4.2.2 Stage 2 results

We again summed participants' scores for each item on each factor and subtracted 4, yielding a score from 0 to 16 on each factor. This time, the scores represented which task the participant believed was better described by each factor. So a participant who scored 0 on the engagement factor strongly believed that the 'Factors' task was more engaging than the 'Product' task, and a participant who scored 16 on this factor believed the reverse. The histograms and descriptive statistics for each factor are shown in Fig. 4.

4.3 Discussion

In both stages of this study, we found substantial between-teacher disagreement about how our tasks were rated on the engagement, demand, routineness and inquiry factors (SDs all ≥2.9). In the case of engagement, in both stages of the study, there were teachers who fell onto every point of our 17-point scale (0–16). In other words, there were some teachers who felt that the 'Trapped Squares' task was as engaging as it is possible to be, and there were some who felt it was as unengaging as it is possible to be. A similar result was found for the demand factor. Consistent with this, in the second stage, we found substantial disagreement about which of the 'Factors' and 'Product' tasks was more engaging or more demanding or more inquiry-based (SDs ≥ 3.2). In contrast, we found more agreement concerning the context factor (SDs of 2.6 and 1.7), presumably because all three tasks randomly selected were situated within the pure domain of number theory. Routineness was also interpreted quite consistently in the second stage (SD = 2.1).

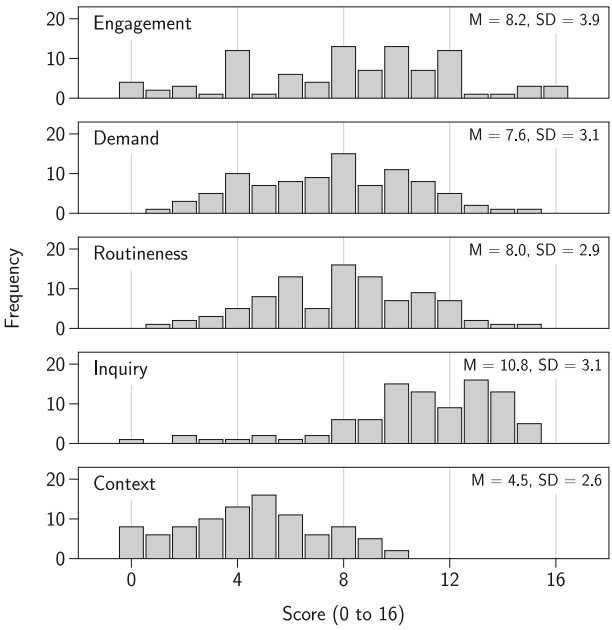

Fig. 3 Histograms showing the distribution of teachers' appraisals of the 'Trapped Squares' task. *M* mean, SD standard deviation

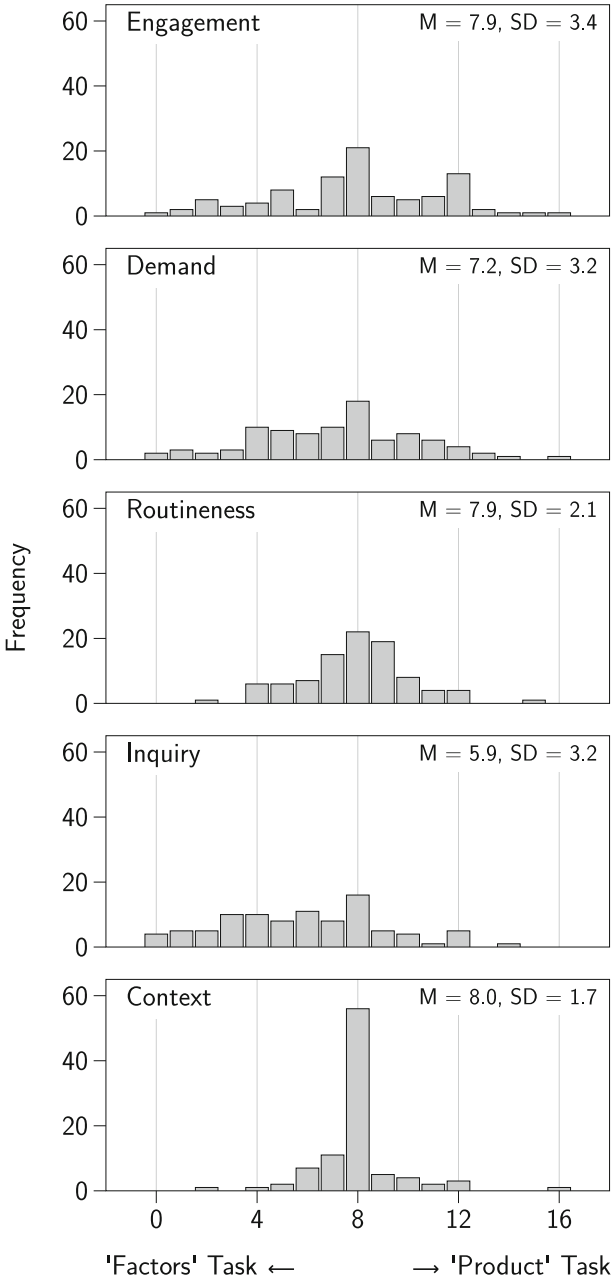

Fig. 4 Histograms showing the distribution of responses to the task in stage 2. Low scores indicate that the participant believed that the 'Factors' task was better described by that factor; high scores indicate that the participant believed that the 'Product' task was. *M* mean, SD standard deviation

5 General discussion

Our goal in this paper was to explore teachers' appraisals of adjectives relating to mathematics tasks and determine whether it was sensible to describe what we have called the task appraisal space in terms of a dimensional structure. In study 1, we found that it was and that the space varied along

seven dimensions, which we termed engagement, demand, routineness, strangeness, inquiry, context and interactivity. The factors were largely orthogonal, with only the correlation between the engagement and inquiry factors exceeding $r = .3$. In study 2, we investigated the extent to which teachers agreed with one another about the properties of specific mathematics tasks. In the first stage, we found that there was between-teacher agreement concerning where the task should be placed on the context dimension, but little agreement on the others. In the second stage, we also found little agreement about which task should be placed higher on the engagement, demand and inquiry dimensions, but found some degree of agreement regarding the routineness and context dimensions.

The findings reported in this paper have two main implications for our understanding of teachers' appraisals of mathematics tasks. First, we discuss the implications of our results for our understanding of the task appraisal space. Second, we discuss implications for the nature of teachers' mathematics task appraisals.

This analysis of the structure of teachers' task appraisals yields several insights. In particular, it sheds light on the sometimes conflicting characterisations of adjectives that have appeared in the literature when describing mathematics tasks. For instance, many task designers have offered descriptions of what they mean by a 'rich' mathematics task (e.g., Ahmed, 1987; Swan, 2008), which seem to include a large array of different characteristics. Although these characterisations share some features, they are not identical and appear to encompass many different properties. Our EFA supports the suggestion that a task's richness does indeed depend on at least two largely independent properties. 'Rich' loaded particularly strongly onto both the inquiry and the engagement factors, with loadings of .59 and .28, respectively, suggesting that richness is a multidimensional notion.

Our analyses also allow us to highlight the independence of task properties which some authors have suggested are related. For example, we found that, contrary to some suggestions in the popular (e.g., Kitchen, 2010) and academic (e.g., Schiefele & Csikszentmihalyi, 1995; Trafton, Reys & Wasman, 2001) literature, engagement and inquiry are perceived by teachers to be only weakly related.

A second implication of our study concerns the nature of teachers' mathematics task appraisals. Our data suggest that these appraisals are reliable in one sense, but largely unreliable in another. We found a large degree of within-teacher consistency: the Cronbach's alpha internal reliability coefficients in study 2 (Table 5, calculated separately for each factor) were consistently high on three of the five factors we considered and not too far from satisfactory on the other two. In other words, if a teacher felt that a given task was 'appealing', they were also extremely likely to believe that it was 'pleasing' (two words from the same 'engaging' factor). Similarly, if a teacher felt that a task was not 'appealing', they were also extremely likely to believe that it was not 'pleasing'. Likewise, responses to the adjectives 'routine', 'procedural' and 'mechanical' were all highly correlated. So we can conclude that teachers' task appraisals are reasonably internally consistent. However, study 2 suggested that teachers' task appraisals show very little *between-teacher* agreement. In stage 1 of study 2, we found that whereas some teachers rated our task as being highly routine, others rated it as being highly non-routine. Similarly, some teachers rated the task as demanding, others as non-demanding; some as inquiry-based, others as non-inquiry-based; and some as engaging, others as non-engaging. In stage 2, we found that this inconsistency also seemed to be present when teachers were asked to engage in *relative* judgement. In short, our results appear inconsistent with the proposal that teachers always agree about the meaning of adjectives that are commonly used to describe properties of mathematics tasks (or at least how they should be assigned to tasks).

Of course, one important criticism which can be levelled at these results is that our data from study 2 pertain only to three specific tasks. Perhaps if we repeated these studies with

different mathematics tasks we would find substantially higher levels of between-teacher agreement. Clearly, replications with other tasks, in other areas, for other ages of students, would be valuable. Ideally, we would randomly sample from the population of mathematics tasks (a difficult endeavour); without this, robustly generalising across all mathematics tasks is impossible. It would also be valuable to explore appraisals of adjectives relating to mathematics tasks by teachers in different cultural and geographical locations. Furthermore, although we have explored how teachers talk about tasks, we have not examined what they *do* with these tasks in their classrooms. Investigating this could help to explain why they appraise tasks in the ways that they do. However, according to the results of study 2, we cannot assume a priori that teachers will agree whether or not a given adjective can properly be applied to a given task, and this has potentially serious implications for practice, an issue we turn to next.

If communication between two people is to be effective, they must have some shared understanding of the meaning of words that form the basis of their communication (Yeo, 2017). Teachers are regularly presented with advice on choosing the tasks that they use in their mathematics lessons, and this guidance often uses some of the adjectives discussed above, such as 'rich' (e.g., DfE, 2014). The danger highlighted by our findings is that teachers may have different understandings of such adjectives from those of the authors, meaning that the curriculum implemented will differ markedly from that intended (Remillard & Bryans, 2004; Remillard & Heck, 2014). Would the 'Trapped Squares' task shown in Fig. 1 satisfy the curriculum authors' 'rich' criterion? Whilst we found that many teachers found the task to be engaging and inquiry-based (the signature characteristics of 'rich' tasks according to study 1), we also found that many regarded it as neither engaging nor inquiry-based.

Might these differences be accounted for by suggesting that our participants were thinking of the task in the context of very different classes? Despite our instruction that the task was aimed at a "typical" class of 13- to 15-year-olds, perhaps some participants were thinking of how a class of high-attaining students might respond to it, whereas others were thinking of how a class of low-attaining students might? We make two comments on this account. First, in order for it to be consistent with the results from the second stage of study 2, one would need to hypothesise that there are some classes for which the 'Product' task is more engaging than the 'Factors' task and others for which the reverse is true. In other words, one would need to propose a relationship between whatever factor differs between the classes (prior attainment, say) and how engaging the class finds the activity. Whilst it does seem possible to concoct examples of activities which would have this property—an extremely difficult mathematics activity might be engaging for a class with high prior attainment and unengaging for a class with low prior attainment, whereas an extremely easy activity might have the reverse properties—the tasks we used in study 2 do not appear to have these properties. Second, if one did wish to argue that the adjectives commonly applied to mathematics tasks can only properly be applied to mathematics activities, then this would seem to have serious implications for the way current advice to teachers is presented. For instance, the English National Curriculum advises that teachers should choose 'rich' problems. If richness is a property of activities and not tasks (Mason, 2015), then the focus of curriculum documents such as this should not be on the tasks chosen by teachers but rather on the way that they are enacted in the classroom. However, even in this case, it would still be necessary for teachers to have *some* basis for selecting one task rather than another.

If, alternatively, one does accept that adjectives such as 'rich' can be properly applied to tasks rather than just activities, then the results of study 2 seem to imply that teachers simply disagree about when tasks have certain properties, at least some of which are commonly discussed in curriculum documents. Teachers' differing perspectives (Philipp, 2007; Simon & Tzur, 1999) are

likely to lead to differing experiences for students in their classrooms. One solution would be for such documents to define such terms carefully using other adjectives. But would there be a shared understanding of these other adjectives? Researchers should be cautious about introducing any new adjectives that might be thought necessary for describing mathematical tasks. Instead, following Burkhardt (2009), we suggest that there should be a much greater role for concrete examples in discussions of task choice. Rather than merely stating that teachers should choose 'rich' tasks, curricula should include detailed examples of the kinds of tasks that the authors have in mind, in the hope that such concrete examples will provide a grounding for adjectives such as 'rich'. Our results suggest that without efforts of this kind, it may be misguided to assume that teachers will interpret discussions about mathematics tasks in the manner intended. Indeed, the adjectives used to describe mathematics tasks may have unexpected interpretations, and this could be one reason why reforms in education sometimes fail.

Acknowledgements This work was partly funded by a Royal Society Worshipful Company of Actuaries Research Fellowship (to MI). We would like to thank the anonymous reviewers for extremely helpful comments on previous versions of this paper.

Appendix

Table 6 Loadings for each adjective on each factor (pattern matrix)

	Engagement	Demand	Routineness	Strangeness	Inquiry	Context	Interactivity
enjoyable	**.81**	−.07	.06	−.02	−.02	−.05	.15
fun	**.74**	−.05	−.05	.21	−.17	.05	.18
pleasing	**.73**	.06	.12	.09	−.02	−.01	.02
appealing	**.68**	−.11	−.04	.09	.02	.02	.07
attention-grabbing	**.68**	.02	−.10	.16	.07	.10	.05
motivating	**.66**	.04	.05	−.11	.08	−.03	.14
stimulating	**.64**	.10	−.02	−.05	.19	−.03	.09
memorable	**.62**	.12	−.01	.06	.04	.19	−.02
boring	**−.58**	.02	.35	.24	.03	−.03	.05
interesting	**.57**	.12	−.04	−.04	.12	.09	.15
absorbing	**.57**	.12	.00	−.10	.18	−.07	.12
exciting	**.57**	−.11	.02	.04	.05	.09	.12
inspiring	**.56**	.08	.08	.00	.19	.06	.10
dull	**−.56**	−.02	.36	.23	.04	−.05	.00
engaging	**.56**	.02	.01	−.11	.10	−.03	.20
amusing	.45	−.10	.03	.44	−.03	.11	−.02
imaginative	.42	−.02	−.15	.24	.29	.09	.05
tedious	−.39	.02	.37	.20	.04	−.03	.01
creative	.39	−.05	−.13	.24	.36	.11	.05
original	.37	.03	−.16	.23	.14	.09	.01
unusual	.35	.12	−.15	.31	.21	.08	−.12
accessible	.28	−.02	.02	−.05	.10	.14	.19
difficult	.12	**.70**	.00	−.07	.07	.00	−.08
complicated	−.02	**.64**	.01	.02	−.06	.07	.05
demanding	.14	**.64**	.08	−.16	.18	.06	−.04
perplexing	.05	**.59**	−.05	.33	.14	−.03	−.08
easy	−.01	**−.58**	.28	.26	.00	.02	−.01
challenging	.24	**.56**	.00	−.24	.12	.02	.00
simple	.06	**−.55**	.19	.28	.07	−.07	−.06
problematic	−.02	**.55**	.01	.26	−.02	.04	−.06

Table 6 (continued)

	Engagement	Demand	Routineness	Strangeness	Inquiry	Context	Interactivity
puzzling	.09	**.52**	−.09	.13	.20	−.14	.04
confusing	−.22	.45	.01	.35	.04	.03	−.07
testing	.29	.42	.17	−.14	−.09	.12	−.09
time-consuming	−.10	.39	.14	.13	.01	−.04	.20
quick	.13	−.32	.17	.10	−.03	.03	−.13
routine	−.09	−.01	**.64**	−.02	−.20	.03	−.14
repetitive	−.05	−.06	**.62**	.07	−.12	−.10	.09
procedural	.10	.01	**.61**	−.04	−.25	.08	.05
formal	−.13	.15	**.60**	−.14	.10	−.01	−.15
mechanical	.01	−.06	**.60**	−.01	−.06	.08	−.10
rule-based	.14	−.04	**.55**	−.04	−.04	−.03	.01
predictable	−.18	−.18	.49	−.04	.00	.06	.02
technical	−.03	.28	.35	−.07	.04	.21	.04
shallow	−.33	−.11	.34	.33	−.13	−.02	−.02
slow	−.23	.13	.33	.31	.00	−.04	.09
passive	−.18	−.09	.33	.20	−.03	.12	−.24
concrete	.15	−.16	.32	−.09	.18	.20	.13
strange	.05	.29	.00	**.54**	.13	−.05	−.20
informal	.18	−.12	−.16	.47	−.06	−.02	.13
irrelevant	−.25	.03	.16	.44	.07	−.22	−.13
serious	−.03	.11	.26	−.38	.30	−.01	−.02
distracting	.06	.07	.23	.34	.06	−.03	−.06
useless	−.14	−.10	.24	.32	.08	−.25	−.15
open	−.01	−.11	−.16	.09	**.69**	.07	.04
inquiry-based	−.08	−.05	−.11	.02	**.67**	.09	.20
deep	.12	.14	−.01	−.09	**.65**	.04	−.04
exploratory	.08	.02	−.07	.06	**.60**	.05	.22
investigative	−.05	.04	−.14	.06	**.60**	.09	.14
rich	.28	.07	−.08	−.04	**.59**	.05	−.02
thought-provoking	.18	.15	−.06	−.01	**.54**	.04	.13
closed	.08	.13	.38	.03	**−.51**	−.04	−.10
analytical	−.03	.20	.06	−.16	**.51**	.04	.05
conceptual	.07	.00	.05	−.13	.49	.02	.07
intriguing	.31	.08	−.19	.14	.46	.06	−.13
experimental	.05	.07	.00	.27	.43	.12	.18
surprising	.35	.00	−.05	.34	.38	.05	−.13
abstract	.07	.26	.14	.10	.37	−.21	−.21
problem-solving	.04	.24	−.10	.00	.31	.25	.10
extendable	.21	.05	.10	−.14	.27	.08	.07
ICT-based	−.06	−.11	.00	.13	.22	.20	−.04
real-life	−.07	.01	.00	.10	−.05	**.86**	−.04
realistic	.05	.03	.04	−.09	−.09	**.72**	−.04
context-based	−.07	.05	−.02	.01	.01	**.66**	−.06
applied	−.04	.05	−.03	.04	−.02	**.64**	.19
authentic	.18	−.01	.04	−.08	.12	.44	−.06
modelling	.02	−.03	.11	.09	.19	.38	.07
meaningful	.21	.05	.05	−.27	.15	.32	.25
multistep	−.02	.23	.15	−.13	.21	.24	.00
hands-on	.09	−.09	.05	.10	.08	.05	**.58**
cooperative	.10	.15	−.02	−.06	.09	.00	**.57**
collaborative	.11	.16	−.05	.02	.08	−.05	**.56**
practical	−.09	−.11	−.03	.11	.04	.39	**.51**
active	.27	−.04	.06	.00	.04	−.03	.48
useful	.23	.01	.06	−.28	.01	.30	.32

Loadings with magnitude >.5 are shown in bold

References

Ahmed, A. (1987). *Better mathematics: A curriculum development study based on the low attainers in mathematics project*. London: HM Stationery Office.

Bräunling, K., & Eichler, A. (2015). Teachers' beliefs systems referring to the teaching and learning of arithmetic. In C. Bernack-Schüler, R. Erens, T. Leuders, & A. Eichler (Eds.), *Views and beliefs in mathematics education: Results of the 19th MAVI Conference* (pp. 95–107). Fachmedien Wiesbaden: Springer.

Burkhardt, H. (2009). On strategic design. *Educational Designer*, *1*(3). Retrieved from http://www.educationaldesigner.org/ed/volume1/issue3/article9.

Christiansen, B., & Walther, G. (1986). Task and activity. In B. Christiansen, A. G. Howson, & M. Otte (Eds.), *Perspectives in mathematics education* (pp. 243–307). Dordrecht: Reidel.

Clarke, D. J. (2013). Contingent conceptions of accomplished practice: The cultural specificity of discourse in and about the mathematics classroom. *ZDM Mathematics Education, 45*(1), 21–33.

Clarke, D. J., & Mesiti, C. (2010). The Lexicon Project: Accessing the pedagogical vocabulary in languages other than English. In M. M. F. Pinto & T. F. Kawasaki (Eds.), *Proceedings of the 34th Conference of the International Group for the Psychology in Mathematics Education* (Vol. 1, pp. 237–238). Belo Horizonte, Brazil: PME.

Cohen, R. J., & Swerdlik, M. (2009). *Psychological testing and assessment: An introduction to tests and measurement* (7th ed.). New York: McGraw-Hill.

Common Core State Standards Initiative (CCSSI). (2010). *Common core state standards for mathematics*. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.

Cuban, L. (1993). The lure of curricular reform and its pitiful history. *The Phi Delta Kappan, 75*(2), 182–185.

Department for Education (DfE). (2014). *Mathematics programmes of study: Key stage 4. National Curriculum in England*. London: Department for Education.

Dziuban, C. D., & Shirkey, E. C. (1974). On the psychometric assessment of correlation matrices. *American Educational Research Journal, 11*(2), 211–216.

Foster, C. (2008). *50 Mathematics lessons: Rich and engaging ideas for secondary mathematics*. London: Continuum.

Foster, C. (2010). *Resources for teaching mathematics 14–16*. London: Continuum.

Foster, C. (2011). *Resources for teaching mathematics 11–14*. London: Continuum.

Foster, C. (2013). *The essential guide to secondary mathematics: Successful and enjoyable teaching and learning*. London: Routledge.

Griffin, P. (2009). What makes a rich task? *Mathematics Teaching, 212*, 32–34.

Gullvåg, I. (1955). Criteria of meaning and analysis of usage. *Synthese, 9*, 341–361.

Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179–185.

Jones, K., & Pepin, B. (2016). Research on mathematics teachers as partners in task design. *Journal of Mathematics Teacher Education, 19*(2–3), 105–121.

Kitchen, M. (2010). *Real-life math: Volume 1: Grades 4–9*. Cuyahoga Falls, OH: CreateSpace Independent Publishing Platform.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral Psychology, 2*, 441–448.

Lengnink, K., & Prediger, S. (2003). Development of personal constructs about mathematical tasks—A qualitative study using repertory grid methodology. In N. A. Pateman, B. J. Dougherty, & J. T. Zilliox (Eds.), *Proceedings of the 27th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 39–46). Honolulu, HI: PME.

Mason, J. (2015). Being mathematical—With, and in-front-of, learners. *Mathematics Teaching, 248*, 15–20.

Mason, J. (2016). When is a problem…? "When" is actually the problem! In P. Felmer, E. Pehkonen, & J. Kilpatrick (Eds.), *Posing and solving mathematical problems* (pp. 263–285). Switzerland: Springer International.

McDonald, S., & Watson, A. (2010). *What's in a task? Generating mathematically rich activity*. London: Qualifications and Curriculum Development Agency.

Næss, A. (1938). Common-sense and truth. *Theoria, 4*, 39–58.
Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language, 24*(5), 586–604.
Philipp, R. A. (2007). Mathematics teachers' beliefs and affect. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 257–315). Charlotte, NC: Information Age Publishing.
Remillard, J. T., & Bryans, M. B. (2004). Teachers' orientations toward mathematics curriculum materials: Implications for teacher learning. *Journal for Research in Mathematics Education*, *35*(5), 352–388.
Remillard, J. T., & Heck, D. J. (2014). Conceptualizing the curriculum enactment process in mathematics education. *ZDM Mathematics Education, 46*(5), 705–718.
Schiefele, U., & Csikszentmihalyi, M. (1995). Motivation and ability as factors in mathematics experience and achievement. *Journal for Research in Mathematics Education, 26*(2), 163–181.
Shimizu, Y., Kaur, B., Huang, R., & Clarke, D. (2010). The role of mathematical tasks in different cultures. In Y. Shimizu, B. Kaur, R. Huang, & D. Clarke (Eds.), *Mathematical tasks in classrooms around the world* (pp. 1–14). Rotterdam: Sense Publishers.
Simon, M. A., & Tzur, R. (1999). Explicating the teacher's perspective from the researchers' perspectives: Generating accounts of mathematics teachers' practice. *Journal for Research in Mathematics Education, 30*, 252–264.
Stylianides, A. J., & Stylianides, G. J. (2008). Studying the classroom implementation of tasks: High-level mathematical tasks embedded in 'real-life' contexts. *Teaching and Teacher Education, 24*, 859–875.
Sullivan, P., Clarke, D., & Clarke, B. (2013). *Teaching with tasks for effective mathematics learning*. New York: Springer.
Swan, M. (2008). A designer speaks: Designing a multiple representation learning experience in secondary algebra. *Educational Designer*, *1*(1). Retrieved from www.educationaldesigner.org/ed/volume1/issue1/article3.
Trafton, P. R., Reys, B. J., & Wasman, D. G. (2001). Standards-based mathematics curriculum materials: A phrase in search of a definition. *Phi Delta Kappan, 83*(3), 259–264.
Watson, A., & Mason, J. (2007). Taken-as-shared: A review of common assumptions about mathematical tasks in teacher education. *Journal of Mathematics Teacher Education, 10*, 205–215.
Watson, A., & Ohtani, M. (Eds.). (2015). *Task design in mathematics education: An ICMI study 22*. Heidelberg: Springer.
Yeo, J. B. W. (2007). *Mathematical tasks: Clarification, classification and choice of suitable tasks for different types of learning and assessment (Tech. Rep. ME2007-01)*. Singapore: National Institute of Education, Nanyang Technological University, Singapore.
Yeo, J. B. (2017). Development of a framework to characterise the openness of mathematical tasks. *International Journal of Science and Mathematics Education, 15*(1), 175–191.