Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 29 Number 10, June 2024 ISSN 1531-7714

Assessing Students' Application Skills Through Contextualized Tasks: Toward a More Comprehensive Framework for Embedding Test Questions in Context

Filio Constantinou, *University of Cambridge*

This study investigated task contextualization as a means of assessing students' ability to apply their subject knowledge to new situations. Through analyzing 527 Functional Mathematics examination questions that claim to assess students' application skills, it developed a set of principles for embedding questions in context: deep contextualization, context balance, context unpredictability and context purposefulness. This paper differentiates between two interpretations of context appropriateness in assessment: (a) the extent to which the context allows students to demonstrate their true knowledge and skills, and (b) the extent to which the context is consistent with the specific aims (or claims) of the course/qualification of which the assessment is part. While the former interpretation has been extensively researched, the latter is less – if at all – explored. This paper examines this latter interpretation. It then combines the two conceptualizations of context appropriateness to propose a more comprehensive framework for assessing students' application skills.

Keywords: contextualized questions; application skills; item writing; assessment design; assessment quality; validity.

Introduction

One of the missions of schooling is to develop students' understanding of different disciplines to prepare them "for the opportunities, responsibilities and experiences of later life" (Department for Education, 2014, p. 4). The understanding that schooling seeks to promote moves beyond the "passive possession of information" and involves "the capacity to act wisely, decisively and effectively […] in context" (Wiggins & McTighe, 2007, p. 13). Evidence that students have developed such understanding is their ability to transfer, or apply, their knowledge to specific situations. Determining, or assessing, the extent to which students have developed this ability is essential: it can inform concerned stakeholders (e.g., teachers, parents, employers) about the level of preparedness of individual students for their next educational or professional step, while providing indications about the effectiveness of schooling in equipping students with the skills required for succeeding in later life.

Gauging, however, students' readiness to apply their knowledge to concrete situations remains a challenge for assessors internationally. One route to addressing this challenge is authentic assessment. Authentic assessment employs authentic tasks, that is, tasks that bear close resemblance to processes normally encountered in real life (see e.g., Koh, 2017; Sokhanvar et al., 2021; Villarroel et al., 2018). Such tasks are typically open-ended, invite a variety of responses, are normally completed over several hours or longer, and tend to take the form of project-based activities (Wiggins, 1989). Arguably, given their realistic and performance-based character, authentic tasks

constitute appropriate tools for measuring students' application skills (see Cumming & Maxwell, 1999).

However, authentic tasks do not always represent the preferred assessment option among assessors, mainly due to the reliability challenges that the evaluation of students' performance in such tasks entails (see e.g., Linn et al., 1991; Olfos & Zulantay, 2007). A tendency to favour more traditional modes of assessment such as written tests, over authentic assessment, is likely to be more common in high-stakes assessment contexts. For instance, in England, A level qualifications, which are normally taken by 18-year-old students and are used for admission to university, have been reformed and are now "in principle exam-only" (Long, 2017, p. 4). Internal assessment that used to take the form of projects and practical activities has been substantially reduced, resulting in written tests becoming the primary and, in most cases, the sole mode of assessment in these qualifications. The aim of this study was to investigate how students' application skills can be assessed through conventional written tests. The study focused specifically on how contexts (scenarios) are used in such tests and sought to develop a framework, or a set of principles, to guide the process of embedding written tasks (questions) in context. The research was carried out as part of informing the reform of Functional Skills in England. Functional Skills are qualifications that aim to promote and assess students' practical skills in English, mathematics and ICT (Information and Communications Technology). Their reform provided an opportunity for revisiting the issue of contextualization in written tests and further reflecting on it.

The Profile of Functional Skills

"Communication" and the "application of number", which are often referred to as "key skills" or "core skills", are typically viewed as essential preparation for employment (Kelly, 2001; OECD, 2019; World Bank, 2023). In 1996, the Dearing Report, a review of higher education in the UK, highlighted employers' concerns about the insufficient levels of these skills among graduates (Dearing, 1996). In response to these concerns, the English government introduced Key Skills, a qualification which aimed to raise the literacy and numeracy standards in the adult population through emphasizing the effective application of English and mathematics knowledge. The introduction of Key Skills was followed by the development of a series of other qualifications, such as Basic Skills and Functional Skills, which claimed to serve a similar purpose. Functional Skills were introduced in 2009 and replaced Key Skills. Their aim was to prepare students for the communicative and mathematical demands of everyday life and the workplace (Education and Training Foundation [ETF], 2015).

In recent years, a review identified the need for reforming Functional Skills to improve their relevance, rigour and credibility in the labour market (ETF, 2015). In response to this review, the English government decided that Functional Skills should be redeveloped. This study was carried out with the aim of informing the redevelopment process, especially that pertaining to the mathematics component of the qualification (henceforth Functional Mathematics).

The most distinctive feature of Functional Mathematics is the emphasis it places on the application of knowledge in everyday-life and workplace settings. In light of this, this study sought to investigate the settings, or contexts, used in Functional Mathematics assessments, with the aim of identifying the type of contextualization that would be appropriate for such assessments. Contextualization here signifies the process of embedding a written question in a scenario.

Contextualizing mathematics tasks: a pedagogical and a psychometric perspective

In mathematics education, the notion of context can refer to either the situation in which a mathematics task is embedded ("task context") or the broader learning environment in which the task is taught ("pedagogical context") (Sullivan et al., 2003). Task context, that is, the focus of this study, may be non-mathematical or mathematical. The wide range of situations in which mathematics tasks can be embedded is captured, for example, in the mathematics framework developed by the Programme for International Student Assessment (PISA). The PISA 2022 framework identifies four types of context in which mathematics assessment tasks can be embedded: personal (e.g., food preparation, travel, shopping), occupational (e.g., costing and ordering of

construction materials), societal (e.g., public transport, voting system, national statistics) and scientific (e.g., climate, medicine, mathematics) (OECD, 2023). Of these, the contexts which are non-mathematical are expected to be realistic and relevant to students' lives. However, this is not always easy to achieve as what is relevant to one student might not be relevant to another (Sullivan et al., 2003). The large heterogeneity that characterizes the student body "individualizes" the nature of the interaction between the students and the context (Boaler, 1993, p. 16), requiring teachers and test developers to exhibit sensitivity and caution when situating tasks in non-mathematical contexts.

The use of non-mathematical context as a means of transforming a mathematics task into a "real-world task" can take a number of different forms and tends to be approached differently by educational systems across the world (see Smith & Morgan, 2016). Real-world tasks can take the form of simplified word problems that contain minimal extra-mathematical information (Beswick, 2011). Such tasks tend to clearly signal the mathematical procedures required for solving the problem, while making use of "camouflage context" (De Lange, 1995): they require minimal engagement with the context and often involve students relying on key words such as "more" and "less" to identify, or infer, the mathematical operations needed for solving the task (Nesher & Teubal, 1975). Real-world tasks can also take the form of problems that simulate those encountered in real life and "for which there are no ready-made algorithms" (Kramarski et al., 2002, p. 226). Such tasks tend to be more authentic in nature, resembling "meaningful, purposeful [and] goal-directed" activities (Jurdak, 2006, p. 284).

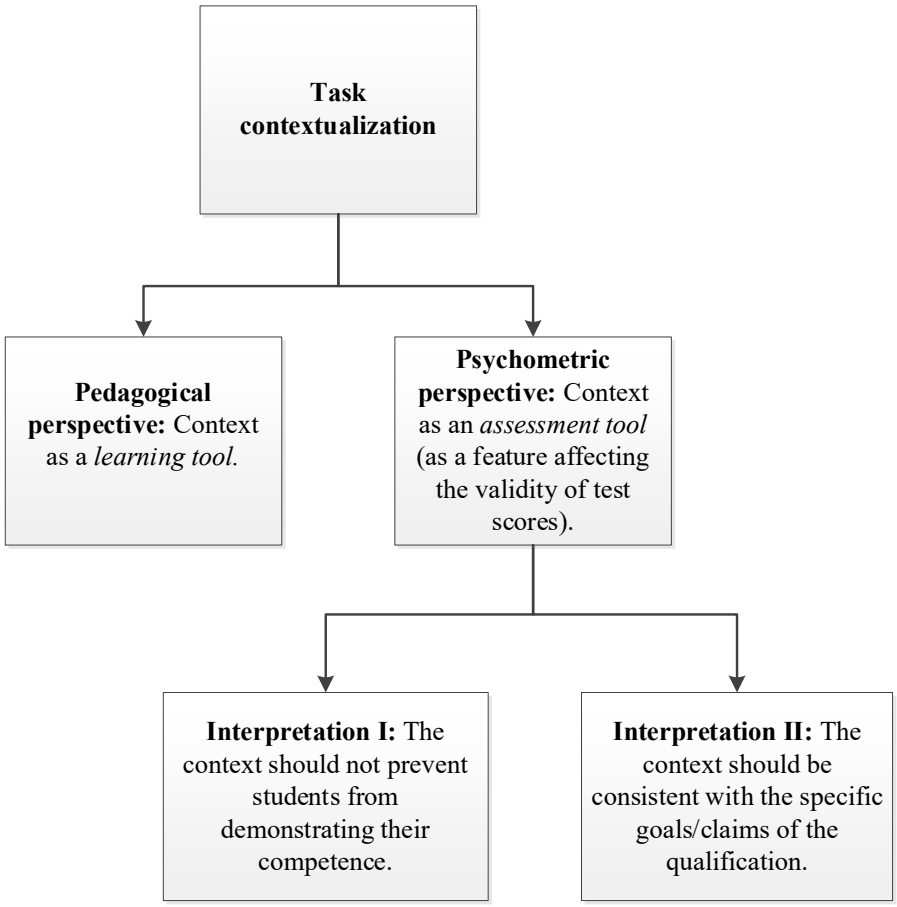
In the literature, the contextualization of mathematics tasks is approached mainly from two perspectives: a pedagogical and a psychometric one (see Figure 1). From a *pedagogical* perspective, contextualization is viewed as a tool for supporting the development of students' mathematical thinking, or students' ability to "mathematize" (see e.g., Gravemeijer & Doorman, 1999). At the heart of this perspective lies the Realistic Mathematics Education (RME) pedagogy which was originally developed by the Freudenthal Institute in the Netherlands (see van den Heuvel-Panhuizen & Drijvers, 2014; van den Heuvel-Panhuizen, 2019). The RME approach advocates the use of context-embedded tasks as a means of inviting students to engage in mathematical modelling. Mathematical modelling, which involves transforming contextualized problems into mathematical problems, is seen to enable students to develop an in-depth understanding of mathematical concepts (Doorman, 2019; Wijaya et al., 2015).

A *psychometric* view of contextualized mathematics tasks, on the other hand, focuses on the assessment rather than the learning of mathematics. In the field of assessment, contextualized mathematics tasks are used as a tool for measuring students' mathematical competence. As a result, most assessment research in this area has examined the ways in which different features of contextualized tasks may interfere with the measurement process. Visual resources, which are often part of contextualized tasks (e.g., Crisp & Sweiry, 2006), and context familiarity (e.g., Boaler, 1994) are two such features that have been investigated with respect to their impact on students' test performance. At the heart of this research lies the notion of validity, defined as the appropriateness of the inferences drawn from students' test scores (see Messick, 1989). However, for appropriate inferences to be drawn from students' test scores, the features of contextualized tasks need to meet two conditions: (a) they should not prevent students from demonstrating their true mathematical knowledge and skills (see Interpretation I, Figure 1), and (b) they should be consistent with the specific aims (or claims) of the course, or qualification, of which the assessment is part (see Interpretation II, Figure 1). Despite its importance, this latter interpretation of context appropriateness does not seem to have been adequately – if at all – researched. To further illuminate this less acknowledged dimension, this study examined Functional Mathematics contextualized tasks in relation to the mission that the Functional Mathematics qualification as a whole claimed to fulfil.

Methodology

The study's methodological approach is presented in detail below. It involves a description of: (a) the process that was followed to define, or operationalize, "context appropriateness" in Functional Mathematics; (b) the assessment materials analyzed; and (c) the process of data analysis.

Figure 1. Task contextualization from a pedagogical and a psychometric perspective



Operationalization of context appropriateness in Functional Mathematics

In line with Interpretation II, this study defined context appropriateness as the extent to which the context used in Functional Mathematics tests was consistent with the aims, or claims, of the qualification. According to the Office of Qualifications and Examinations Regulation (Ofqual) (2015), the overall aim of Functional Skills tests was to "allow students to demonstrate that they have achieved *practical skills* in literacy, numeracy and IT [Information Technology] that help them to live and work confidently, effectively and independently" (author's emphasis) (p. 5). In Functional Mathematics, this aim translated into a "focus on *functionality* and the effective *application* of process skills in *purposeful* contexts and scenarios that reflect *real-life situations*" (author's emphasis) (Ofqual, 2011, p. 9). As it becomes evident from the "manifesto" of the qualification (i.e., the purpose of the qualification as it is articulated in governmental documents[1]), the most salient characteristics of Functional Mathematics were arguably application and purposefulness. Application subsumed concepts such as "practical skills" and "functionality", while purposefulness encompassed the idea that contexts should "reflect real-life situations". Based on these observations, this study defined context appropriateness as the degree to which the context used in Functional Mathematics tests fulfilled the interrelated criteria of application and purposefulness. These two criteria are further analyzed below.

[1] In England, the purpose and subject content of qualifications taken by school-aged students (e.g., GCSEs, A levels) are normally determined by the government. This information is typically captured in various official documents published by the Department for Education (DfE) and the Office of Qualifications and Examinations Regulation (Ofqual) (see e.g., Ofqual, 2011, 2015).

Criterion 1: Application. As a cognitive process, application refers to the "use of abstractions in particular and concrete situations" (Bloom et al., 1956, p. 205). Application constitutes the third level of Bloom's taxonomy, a hierarchical classification system of processes of thinking and learning. According to Bloom's taxonomy, "applying" (third level) is a more cognitively demanding skill than "understanding" (second level) and "remembering" (first level), suggesting that using knowledge to respond to new situations is a process that involves a higher degree of complexity than merely retrieving learnt information (Krathwohl, 2002). Informed by Bloom's taxonomy, the present study deemed the criterion of application to be fulfilled when the context used in the tests invited the students to apply their mathematical knowledge to a new, concrete situation.

Criterion 2: Purposefulness. The criterion of purposefulness was considered to be satisfied when the contexts used in the Functional Mathematics tests were meaningful for the students. Based on the manifesto of Functional Skills as presented above, meaningful contexts can be understood as those contexts that are encountered in everyday life and the workplace. To determine, however, how meaningful, or purposeful, a context is, it is important to first understand how mathematics is normally used in everyday life and the workplace.

In the literature, the mathematics knowledge and skills used in everyday life and the workplace have been described as "culturally determined, socially distributed, technology reliant, technologically embedded, contextual, local, personally invented, and basic." (Gainsburg, 2005, p. 9). The term "basic" is key in this description; the level of mathematics required for coping with the demands of everyday life and the workplace is elementary and corresponds to that covered in the GCSE[2] curriculum (Hodgen & Marks, 2013). However, the contexts in which this basic mathematics is used are typically sophisticated (Steen, 2003) and messy (Wake & Williams, 2003). In fact, the mathematics used in the workplace has been described as "simple mathematics in complex settings", a feature that renders the transfer of school mathematics to the workplace a complicated process (Hodgen & Marks, 2013).

Examples of simple mathematics used in complex workplace contexts include: nurses calculating drug dosages (Hoyles et al., 2001); shop assistants estimating the number of replacement products required to fill the space on the shelves (Hastwell et al., 2013); lab technicians measuring the concentration of potassium in drinking water at a water bottling plant; hairdressers calculating the amount of hydrogen peroxide needed for bleaching their clients' hair; and bakers weighing and measuring quantities of ingredients (Bakker et al., 2011).

Against the backdrop of these observations, this study considered the criterion of purposefulness to be fulfilled when the context required students to use their mathematical knowledge in a manner that resembled that normally encountered in everyday life and the workplace.

Sample

This study analyzed the contexts used in 37 Functional Mathematics examination papers. As can be seen in Table 1, overall, the papers consisted of 527 tasks (or questions) and 69 contexts (or overarching scenarios). All papers were obtained from the legacy qualification and were developed by three major providers of Functional Skills qualifications in England. They spanned all levels of difficulty at which Functional Mathematics is available: Entry Level, which comprises Entry Level 1, Entry Level 2 and Entry Level 3; Level 1; and Level 2. In England, there are eight qualification levels (with a doctorate classified as Level 8). According to Ofqual (n.d), Entry Level is below GCSE, Level 1 is comparable to the bottom GCSE grade range (grades 1 to 3), while Level 2 is comparable to the top GCSE grade range (grades 4 to 9). The duration of the assessments leading to a Functional Mathematics qualification varied across examination boards: it ranged from one hour to one hour and thirty minutes at Entry Level, and from one hour and thirty minutes to two hours at Levels 1 and 2.

[2] GCSEs (General Certificate of Secondary Education) are national, academic qualifications taken by 14- to 16-year-olds in England, Wales and Northern Ireland. They are offered in a wide range of subjects including mathematics. They are graded based on a scale that ranges from 1 (lowest grade) to 9 (highest grade).

Table 1. Sample of examination papers

Level	Number of examination papers	Number of tasks (questions)	Number of contexts (overarching scenarios)
Entry Level 1	8	95	8
Entry Level 2	7	83	7
Entry Level 3	6	64	6
Level 1	8	144	24
Level 2	8	141	24
Overall	37	527	69

The sample analyzed comprised both specimen and past examination papers. Specimen papers tend to become the template for live papers (Constantinou et al., 2018) and, therefore, their inclusion in the analysis was considered to be important. The sample was restricted to papers that were available on the examination boards' websites at the time of the research.

Functional Mathematics is a qualification regulated by Ofqual, England's examinations regulator. As such, all Functional Mathematics papers – specimen papers and past papers – may be seen to embody the same (or largely the same) assessment standard. Developing papers that are comparable to one another (i.e., comparable across examination boards as well as across examination sessions) is one means of ensuring that grades awarded by different examination boards and in different examination sessions are comparable (e.g., in England, a grade "A" awarded by examination board X in 2023 is treated as comparable to – or as carrying the same meaning as – a grade "A" awarded by examination board Z in the same year; a grade "A" awarded in 2023 is treated as comparable to a grade "A" awarded in 2022 and so on). Maintaining standards both across examination sessions and across examination boards is an important consideration in the English examination system and constitutes one of Ofqual's key missions.

It should be clarified that only examination papers were analyzed as part of this study. Any instructions, guidelines or other materials that may have been given to question writers to guide or support the assessment development process were not examined.

Analysis

As the aim of the research was to understand the issue of task contextualization in depth and develop a framework of principles, a qualitative approach to the analysis of the examination questions was taken (see e.g., Flick, 2018). The qualitative analysis was carried out by the author of this paper (experienced assessment researcher), and consisted of two stages. The first stage involved a description of the contextualization used in the papers, focusing specifically on the amount and type of scenarios used. This was followed by an evaluation of the extent to which this contextualization fulfilled the two interrelated criteria of context appropriateness specified above, namely, application and purposefulness. This process of evaluation resulted in a series of observations, or overarching themes. The themes were identified in a largely bottom-up manner and, as is typically the case with exploratory qualitative analysis, they reflect the interpretations of the researcher who developed them. Examples of tasks that led to the development of the themes are provided below to render the analysis more transparent and to enable readers to make their own judgement about the appropriateness of the themes and of any interpretations drawn.

It should be noted that the aim of this paper is not to provide an evaluation of the efficacy of the legacy Functional Mathematics qualification. Rather, its intention is to share and discuss a series of qualitative observations to enhance our understanding of task contextualization and encourage further reflection on this issue. As will be seen below, at various points, attempts were made to quantify some of the qualitative observations. These, however, served merely exploratory purposes and were intended to facilitate the illumination of less overt aspects of task contextualization.

Findings

The qualitative analysis led to the identification of four themes which illuminated some strong and weak aspects of contextualization in Functional Mathematics. These themes, which are presented below, formed the basis for the development of the contextualization framework.

Nested approach to contextualization

As the analysis showed, the examination papers exhibited heavy contextualization. In fact, *all* the questions in the papers were embedded in a scenario. Such a high level of contextualization seems to be more consistent with the applied character of the qualification. Interestingly, the type of contextualization found in Functional Mathematics was substantially different from that normally encountered in the "academic" counterpart of Functional Mathematics, namely GCSE Mathematics. While in GCSE Mathematics contextualized questions are embedded in scenarios that are typically thematically unrelated to one another, in Functional Mathematics this was not the case. In Functional Mathematics, the questions comprising individual papers were thematically interconnected: they all revolved around one storyline (see Entry Level papers), or three different storylines (see Level 1 and Level 2 papers).

For example, one of the Entry Level 2 papers analyzed comprised fifteen questions. All of them revolved around the same theme, that of inviting friends for dinner. This theme, in turn, consisted of four sub-themes:

(1) Going into town to buy food for the guests (the questions invited students to calculate the travel time required to reach the town, transport cost, etc.) (see also Table 2).

(2) Buying the food (the questions focused on issues such as the quantity of food items, food cost, paying for the food, etc.).

(3) Cooking the food (the questions concerned cooking time, dividing the food into portions, etc.).

(4) Entertaining the guests (the questions asked students to split the guests into groups to play a game, etc.).

These four sub-themes essentially represent four mini stories. The mini stories are connected with one another and together they form part of a more overarching narrative, that of inviting friends for dinner. This model of contextualization exemplified here can be described as a "nested", or a "scenario-within-a-scenario", approach to contextualization.

This nested approach to contextualization, which was evident in *all* the examination papers analyzed, helped to transform individual examination papers into narratives that unfolded as the students proceeded from one question to the next. The double layer of context observed in the papers, whereby a story was embedded within a story, is a literary device that is commonly found in novels, plays and films. The "deep" form of contextualization that results from this technique is likely to encourage students to immerse themselves in the story described in the paper, rendering the paper almost a "simulation", or a metaphor, of real life. This effect was strengthened by the directive style in which many of the scenarios were formulated and the use of the pronoun "you" (e.g., *"You ask some friends to come to your house… You need to buy some food… Work out the cost of your food items…"*) which invited students to engage in a form of role-play, albeit a more static and passive role-play than that encountered in conventional performance-based role-playing activities. Given that it can render the process of completing the task somewhat more "experiential" in nature, this form of contextualization may be viewed as a more appropriate tool for assessing students' ability to apply their knowledge. More research is, however, needed before any definitive claims can be made.

Context imbalance

The scenarios in which the Functional Mathematics questions were embedded stemmed from everyday life and the workplace. The former included activities such as going on holiday, making a cake, inviting friends for a meal, going to the cinema, visiting a friend who lives abroad, buying fuel for the car, and buying presents for friends. The latter included working at a coffee shop, working at an ironing shop, working at a bread shop, and setting up a childminding business. As can be seen in Table 3, overall, the vast majority of the scenarios used in the papers analyzed were based on everyday-life activities. This was particularly the case for the Entry Level papers, with

Table 2. An example of a sub-theme (Entry Level 2 paper)

Overarching theme (storyline)	You ask some friends to come to your house at 5 o'clock today.
Sub-theme (mini story)	You need to go to town to buy some food.
Questions	1. The clock shows the time now. What is the time?
	2. You can take a bus, train or taxi into town. How will you travel into town? Make a note of the cost. [The students are provided with information about travel times and fares]
	3. How will you travel back home from town? Make a note of the cost.
	4. Give one reason for your choices.
	5. How much will your travel cost altogether?

Table 3. Percentage of everyday-life and workplace scenarios across levels

	Percentage of everyday-life scenarios	**Percentage of workplace scenarios**
Entry Level 1	87.5%	12.5%
Entry Level 2	100.0%	0.0%
Entry Level 3	100.0%	0.0%
Level 1	45.8%	54.2%
Level 2	54.2%	45.8%

the Entry Level 2 and Entry Level 3 ones consisting exclusively of such scenarios. While Level 1 and Level 2 papers appear more balanced, a closer look at the composition of individual Level 1 and Level 2 papers suggests that this may only be superficial. As Table 4 indicates, some Level 1 and Level 2 papers consisted exclusively either of everyday-life scenarios (see Level 1, Paper 4; Level 2, Papers 3 and 7) or of workplace ones (see Level 1, Papers 2 and 3; Level 2, Papers 1 and 8) and, as such, could not be deemed as balanced. This observed imbalance, or overall dominance of everyday-life scenarios over workplace ones, raised questions about the ability of the qualification to fulfil its dual mission which, according to the Functional Skills manifesto, was to prepare students for the mathematical demands of *both* everyday life and the workplace.

Fulfilling such a dual mission constitutes an ambitious goal the achievement of which would probably present a significant challenge for any qualification. This challenge derives mainly from the large heterogeneity that characterizes the workplace relative to everyday life. While everyday life tends to involve more "universal" tasks with which many students are likely to be familiar either directly or indirectly (e.g., inviting friends for a meal, going to the cinema, buying presents), the same cannot be said about the workplace. As discussed earlier, the workplace comprises multiple highly specialized vocational routes each of which features a different set of tasks. Given the diverse nature of the workplace, it is reasonable to assume that not all students will be familiar with the processes and particularities of every single vocational sector. Therefore, avoiding embedding examination questions in work-related scenarios may emerge as an inevitable practice for examination developers. It is a practice that can provide all students with the same opportunity to succeed in the examinations, irrespective of background and professional experience. Hence probably the overall prevalence of everyday-life scenarios over professional ones in the Functional Mathematics examination papers analyzed.

While capable of enhancing to some extent the fairness of the assessment, this observed inconsistency between the proclaimed mission of Functional Mathematics and the content of the assessment is not without negative consequences. In particular, the gap between *claims* (i.e., Functional Mathematics claimed to prepare students for the mathematical demands of *both* everyday life and the workplace) and *assessment* (i.e., the vast majority of Functional Mathematics assessment tasks were based on everyday-life activities) can undermine the validity of the inferences drawn from students' examination scores across the different levels. For instance, influenced by the claims of

Functional Mathematics, prospective users of students' scores (e.g., employers) may interpret a high score in the examination as an indication that the student will be able to effectively apply their mathematical knowledge to *any* professional setting. However, such an interpretation may not be accurate or appropriate, as the Functional Mathematics score may serve mainly as a reflection of students' ability to apply their mathematical knowledge to an everyday-life context. Using knowledge acquired in an everyday-life context to deal with mathematical challenges in a professional context might be possible. However, such transfer of knowledge from one context to another should not be viewed as an automatic process, or as a process that takes place naturally and in an unassisted manner (see e.g., Anderson et al., 1996). Transfer of knowledge from an everyday-life context to a professional context requires a considerable cognitive leap on the part of the students, especially those of low mathematical ability. In the absence of relevant assessment, it is unknown whether and to what extent Functional Mathematics students would be able to make such a cognitive leap.

Context predictability

During the analysis, a certain degree of uniformity was noticed across papers, which raised questions about the predictability of the papers. As Table 5 indicates, many of the overarching scenarios, or themes, in which the examination questions were embedded occurred in more than one examination paper. For example, the themes "selling products" and "going on holiday" each occurred in three out of the eight Level 2 papers analyzed.

Table 4. Balance of everyday-life and workplace scenarios within individual Level 1 and Level 2 papers

		Number of everyday-life scenarios	**Number of workplace scenarios**
Level 1*	Paper 1**	1	2
	Paper 2	0	3
	Paper 3	0	3
	Paper 4	3	0
	Paper 5	2	1
	Paper 6	2	1
	Paper 7	1	2
	Paper 8	2	1
	Overall	11	13
Level 2	Paper 1	0	3
	Paper 2	1	2
	Paper 3	3	0
	Paper 4	2	1
	Paper 5	2	1
	Paper 6	2	1
	Paper 7	3	0
	Paper 8	0	3
	Overall	13	11

Notes:

* Overall, eight Level 1 and eight Level 2 papers were analyzed. Each set consisted of a combination of specimen papers ("blueprint" papers/prototypes on which future papers were modelled) and past papers (papers taken in different years by different cohorts of students).

**Level 1 and Level 2 papers each consisted of three overarching scenarios (or storylines).

Table 5. Recurring themes (number of examination papers containing each theme)

	Number of papers containing each theme		
	Theme 1: Selling products (e.g., potatoes, teddy bears, TV sound bars, cars, food)	**Theme 2: Going on holiday**	**Theme 3: Organizing an event (e.g., bowling game, sports day)**
Entry Level 1 papers (n=8)	3	2	0
Entry Level 2 papers (n=7)	1	3	1
Entry Level 3 papers (n=6)	3	2	1
Level 1 papers (n=8)	2	1	3
Level 2 papers (n=8)	3	3	2
Overall (n=37)	12	11	7

A pattern of recurrence was not only observed in the themes that featured in the examination papers, but also in the structure of the examination questions. The structural homogeneity noticed across papers was explored more systematically through a small-scale mapping exercise between examination questions and their assessment objectives. The mapping exercise, due to its exploratory nature, involved five examination papers which comprised 53 questions in total. The papers were all obtained from the same examination board to allow the observed pattern of recurrence to be investigated more meaningfully. Mapping examination questions onto the assessment objectives that they targeted allowed examination questions to be grouped according to their assessment function. This enabled the isolation of functionally similar questions, which in turn facilitated the identification of any instances of structural repetition across questions.

The mapping exercise revealed that, in a number of cases, questions that targeted the same assessment objective and could therefore be described as functionally similar, displayed also structural similarity. Examples included:

(1) In Paper A, the candidates were presented with pictures of coins and were asked to *"tick the coins he uses to pay for the cake"*. A similar task appeared in Paper B: *"Tick the money you will use to pay for your main meal and your dessert."*

(2) In Paper C, the candidates were asked to *"Compare the money he [the customer] gives you with the cost of a panini and hot drink."*. A similar question was included in Paper D: *"Compare the money you have with the total cost of your presents."*

(3) All five papers included a task which explicitly asked candidates to name the shape of an object:

- *"Name the shape of the base of the tin you chose."*
- *"Tick the box the fish and chips should come in. What is the name of this shape?"*
- *"Choose a box to put the chocolate cake in. Describe the size and shape of the cake box you chose."*
- *"Tick the shape for the red zone. What is the name of this shape?"*
- *"What shape or shapes are the boxes you have chosen?"*

The recurrence of questions that displayed simultaneously functional and structural similarity appeared to "homogenize", to some extent, the examination papers analyzed. A degree of homogeneity across papers is generally desirable: it ensures that students will be sufficiently familiar with the format of the examination and with what they are expected to do, which can reduce test anxiety and enable students to demonstrate their true abilities.

While a certain degree of familiarity with the format of examination papers is generally beneficial, excessive familiarity can be detrimental. This is because excessive familiarity can increase the predictability of examination papers. Predictability can develop when examination questions that exhibit simultaneously functional, structural and thematic similarity occur in a number of examination papers. Two such questions are as follows:

- *"Tick the box the fish and chips should come in. What is the name of this shape?"* (Paper B)

- *"Choose a box to put the chocolate cake in. Describe the size and shape of the cake box you chose."* (Paper C)

These two questions, which appeared in two different papers, (a) assess the same skill (functional similarity), (b) are structured in a similar way (structural similarity), and (c) are embedded in the same theme, that of packaging takeaway food (thematic similarity). These three layers of similarity render the questions formulaic to the extent that familiarity with one question can increase the likelihood of performing well in the other. Such formulaic and, consequently, predictable questions can compromise the application goal of the qualification, as they are likely to measure students' ability to retrieve previously learnt information rather than their capacity to apply their subject knowledge to a new situation. In addition, they encourage students to view the scenario in which a question is embedded simply as an "add-on", or surface, feature of the question (cf. "camouflage context") rather than as an integral part of the task. A view of context as irrelevant information (i.e., as information that should be ignored rather than as data that is necessary for carrying out the task), is problematic as it may prevent the question from measuring what it intended to measure, namely, students' ability to apply their knowledge to a new situation. These risks posed by predictable questions are particularly relevant to, or salient in, England (as well as other countries with a similar examination system). In England, examination papers are typically released after the examination session for transparency (Baird et al., 2014). As past papers are normally publicly available, teachers and students tend to use them as examination preparation material (e.g., for practice or revision purposes). If questions that present simultaneously functional, structural and thematic similarity occur across different examination sessions, they can easily become formulaic lowering the examination's cognitive demand and compromising its application intentions.

However, despite their negative effects, such formulaic questions cannot always be avoided. This is mainly due to the nature of some types of knowledge and skill included in the Functional Mathematics syllabus. In particular, some types of knowledge and skill are so specific and concrete that it is often difficult for examination question writers to assess them in diverse and non-formulaic ways. For example, one of the assessment objectives in Entry Level Functional Mathematics was to recognize and name common shapes. In practice, there might not be many ways in which this objective can be assessed in an examination paper, which can probably explain the strong structural similarity displayed by five of the questions analyzed as part of the mapping exercise. These five questions, each obtained from a different paper, were as follows (also listed earlier):

- *"Name the shape of the base of the tin you chose."*
- *"Tick the box the fish and chips should come in. What is the name of this shape?"*
- *"Choose a box to put the chocolate cake in. Describe the size and shape of the cake box you chose."*
- *"Tick the shape for the red zone. What is the name of this shape?"*
- *"What shape or shapes are the boxes you have chosen?"*

Context purposefulness

According to the manifesto of Functional Mathematics, a key mission of the qualification was to assess students' ability to apply their mathematical knowledge "in purposeful contexts and scenarios that reflect real-life situations" (Ofqual, 2011b, p. 9). As indicated earlier, the scenarios in which the tasks were embedded stemmed from everyday life and the workplace. While association with everyday life and the workplace may be a necessary condition for designing realistic tasks, it is not a sufficient one. A closer inspection of the tasks revealed that many of them could not be classified entirely as realistic. The feature which appeared to detach many of the tasks from real life was the lack of a genuine real-life incentive for students to carry them out. The lack of such an incentive may have rendered the tasks less purposeful for the students. Examples included:

- *"Choose a tin to cook your cake in. Tick the tin you chose. Name the shape of the base of the tin you chose."*
- *"This is a window in one of the houses. What is the name of the 2D shape of window?"*
- *"Which of these bottles is the shortest?"*
- *"One of the bird feeders comes in a box that is a cuboid. How many faces does the box have?"*
- *"Which birds on the sheet are longer than the blackbird?"*

- *"Anna has 234 Euros for her holiday. What is 234 rounded to the nearest 10?"*
- *"John left 9 packets of dog snacks. The dog eats 3 packets. What fraction of the packets does the dog eat?"*
- *"You have these bags. Which bags weigh more than 10 kg?"*

These tasks do not seem to invite students to use mathematics in a meaningful and purposeful way. For instance, it may not be evident to students why it is important to know the number of faces of a box containing a bird feeder, or which bird from a list of birds is longer than the blackbird. These tasks are in contrast with how mathematics is typically used in the real world. Typically, in the real world, mathematics is used for a reason. For example, it is used to estimate how many products should be collected from the stock room to refill the shelves at a supermarket, to calculate the right quantity of flour required for making 15 loaves of bread at a bakery, or to calculate the right dosage of a drug to administer to a patient at the hospital (see e.g., Bakker et al., 2011; Hastwell et al., 2013). The above tasks do appear to be realistic as they make reference to real-world entities (e.g., tin, bag, house window, bottles, birds, dog snacks). However, they are realistic only on the surface, as the only motivation that the students have for answering these questions is that of demonstrating to their teacher or assessor that they are proficient in mathematics.

The format of presentation of these tasks tends to place them more in the category of "school mathematics" rather than in "functional mathematics". Even though in their current form they tend to resemble "school mathematics", following some modifications, they may have the potential to become more suited to the goals of Functional Mathematics. For instance, the last question ("*You have these bags. Which bags weigh more than 10 kg?*"), when changed to *"You have these bags. Only bags weighing up to 10kg are allowed on the plane. Which bag can you take on the plane?"*, could become a more appropriate question for a qualification that claims to assess students' mathematical knowledge in a purposeful manner. However, more research is required before any definitive conclusions can be drawn about the effectiveness of such modifications.

One issue worth noting with respect to such modifications is that embedding a task in a real-world context and demonstrating its purposefulness – as attempted in the example above – may require a larger amount of text. However, a higher amount of text can increase the reading demand of the question. This is generally undesirable, as increasing the reading demand of a question that was not designed to measure reading ability can introduce construct-irrelevant variance into students' examination scores, undermining the validity of the inferences drawn from them (Haladyna & Downing, 2004; Wiliam, 2008). Therefore, purposeful contextualization should be attempted with caution.

Interestingly, in the sample of examination papers analyzed, tasks exhibiting a lower degree of purposefulness and therefore resembling "school mathematics" were more common in the Entry Level papers (see Table 6). The frequency of these tasks seemed to decline as the level – and, by extension, the difficulty – of the paper increased (with Level 2 papers containing the least amount of such tasks). The higher number of tasks exhibiting a lower degree of purposefulness observed in the Entry Level papers (relative to Level 1 and Level 2 papers) could suggest an attempt on the part of examination question writers to reduce the reading demand of Entry Level tasks as these were intended for less numerate and potentially also less literate students (relative to Level 1 and Level 2 students).

Discussion

Developing students' ability to apply their knowledge to a new situation constitutes a key goal of education internationally. To ensure that education is successful at promoting this goal, appropriate methods of assessing students' readiness to apply their knowledge to new situations need to be identified. The present study investigated how this skill can be assessed through a popular assessment tool, namely written tests. It focused specifically on contextualization, or the process of embedding a written question in a scenario. Contextualization represents the most conventional route to assessing application of knowledge in written tests, and is one of the issues that dominated the discussions that surrounded the most recent reform of Functional Mathematics in England. The aim of this research was to develop a principled approach to contextualizing questions.

Table 6. Percentage of tasks exhibiting a lower degree of purposefulness across levels

Level	Percentage of tasks exhibiting a lower degree of purposefulness
Entry Level 1	49.6%
Entry Level 2	43.5%
Entry Level 3	38.9%
Level 1	32.2%
Level 2	16.8%

Investigating contextualization from a psychometric perspective: Interpretations I and II

The investigation of contextualization in this study was informed by important theoretical distinctions which are not typically articulated explicitly in the literature. The first theoretical distinction concerns the two uses of context in education:

(a) context as a learning tool (pedagogical perspective), and

(b) context as an assessment tool (psychometric perspective).

The second theoretical distinction concerns specifically the use of context as an assessment tool (psychometric perspective), and draws attention to two different interpretations of context appropriateness in assessment:

(a) the extent to which the context does not prevent students from demonstrating their true competence (Interpretation I), and

(b) the extent to which the context is consistent with the specific goals (or claims) of the course, or qualification, of which the assessment is part (Interpretation II).

While both Interpretation I and Interpretation II are linked to validity and are concerned with the design of assessments that are fit for purpose, the latter represents an angle from which contextualization has not been researched to date. To address this gap, this study examined contextualization through the lens of Interpretation II.

Assessing application of knowledge: four contextualization principles

In line with Interpretation II, this study defined context appropriateness as the extent to which the contexts, or scenarios, in which Functional Mathematics questions were embedded, were consistent with the claims of the qualification. As mentioned earlier, the qualification's claims were encapsulated in two interrelated concepts that appeared in the manifesto of Functional Mathematics: application and purposefulness. Application referred to the ability of the qualification to assess students' readiness to apply their subject knowledge to new situations, while purposefulness concerned the assessment of students' subject knowledge through meaningful real-life scenarios.

The study identified four fundamental principles for contextualizing questions in Functional Mathematics examination papers. While these principles were developed through research on Functional Mathematics, they are relevant to all qualifications that strive to assess students' practical skills through written tests. However, they should be experimentally investigated and validated before being formally adopted.

Figure 2 below attempts to translate these principles into a framework aimed at supporting assessment practitioners in developing and/or evaluating tests that assess application of knowledge. The framework consists of a series of questions intended to trigger reflection and enable assessment practitioners to approach task contextualization in a more principled, conscious and deliberate manner.

Principle 1: Deep contextualization. Deep contextualization, as achieved through a nested or a scenario-within-a-scenario approach to contextualization, may be a more appropriate tool for assessing students' application skills compared to "lighter" models of contextualization. The double layer of context it involves seems to invite students to immerse themselves in the scenario and engage in a form of role-play. This can render the process of completing the task somewhat more experiential in

nature, thereby serving the application goal of the qualification more effectively. Deep contextualization stands in contrast to common contextualization practice which involves embedding questions in scenarios that are thematically unrelated to one another (see e.g., GCSE Mathematics). Deep contextualization therefore emerges as an alternative model of contextualization, one that may be more suited to qualifications that make stronger claims regarding application of knowledge.

Principle 2: Context balance. Measuring effectively students' ability to apply their knowledge to new situations presupposes a clear understanding of the types of situations to which students are expected to be able to apply their knowledge. These situations are typically specified in the "manifesto" of the qualification. The manifesto, which articulates the overall mission of the programme of study, needs to be consulted before any decisions are made regarding the nature and content of the assessment. When the types of situations that are relevant to the mission of the qualification are clarified, an attempt should be made for these to be represented in the assessment in a balanced manner. Context balance is a perspective that seems to be currently missing from the assessment literature surrounding contextualization.

Principle 3: Context unpredictability. Where possible, context-embedded questions that are similar in structure, theme and function to context-embedded questions occurring in past examination papers, should be avoided. Such questions can become formulaic and, therefore, predictable. As explained earlier, predictability is undesirable as it can subject context-embedded questions to a process of "degeneration": it can reduce them from "apply" tasks (third level of Bloom's taxonomy) to merely "remember" tasks (first level of Bloom's taxonomy), thereby undermining their ability to fulfil their key mission which is to assess application of knowledge. The three components of context-embedded questions identified in this study (i.e., structure, theme and function), coupled with the mapping exercise undertaken, provide a methodological approach for examining predictability while equipping assessors with a practical tool for detecting highly predictable questions.

Principle 4: Context purposefulness. The tasks should be deeply rooted in the real world and not be only superficially linked to it. Genuinely realistic tasks are those that address explicitly the "why" question and provide students with an authentic real-life incentive to engage with them, one that goes beyond demonstrating to a teacher or assessor that they have acquired the target subject knowledge. As argued in this study, tasks that display these characteristics can be described as "purposeful". Purposeful tasks, in turn, are desirable as they can provide opportunities for assessing students' application skills in a more authentic manner.

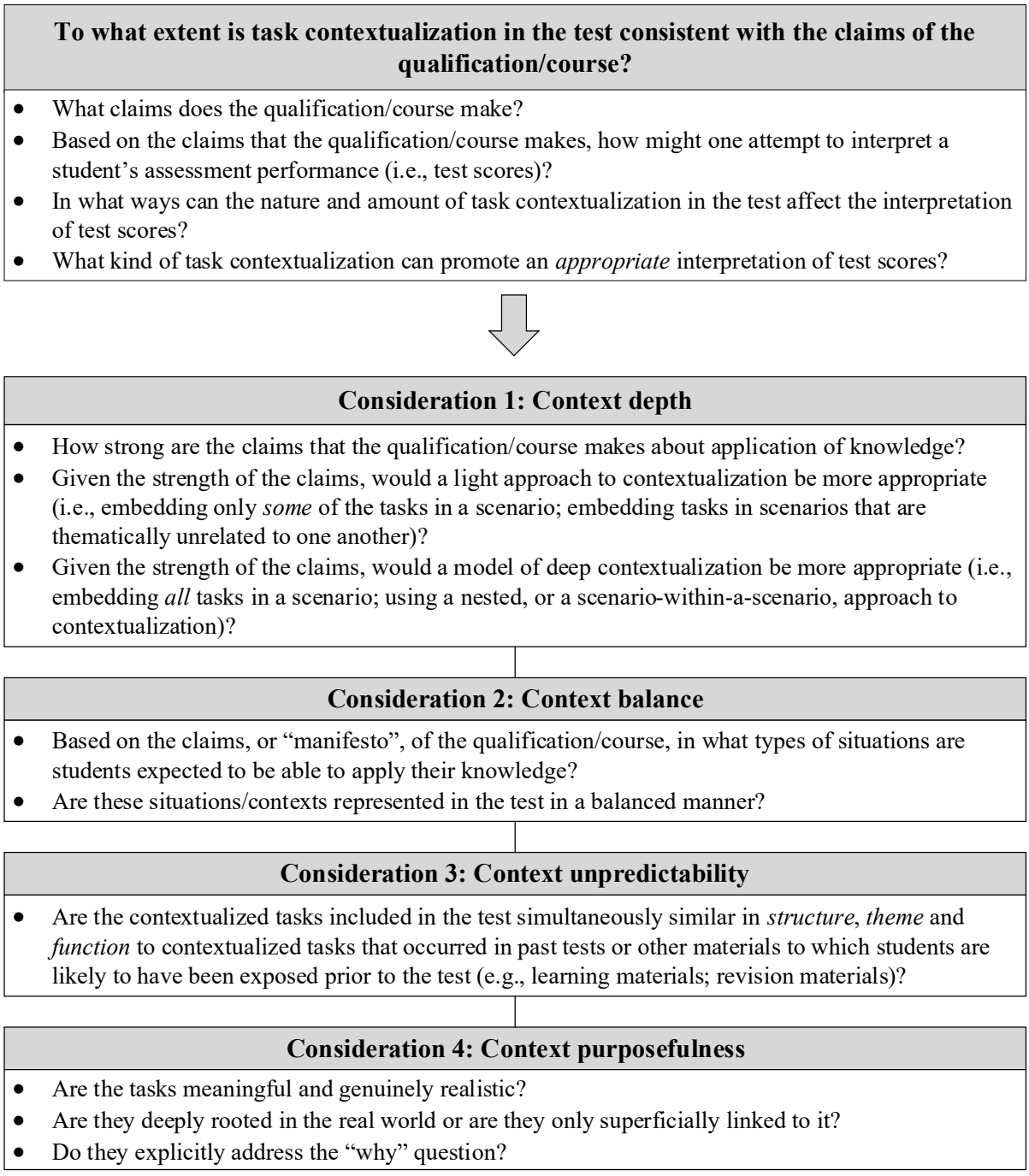
The descriptor "realistic", which is commonly used in the literature on contextualization, does not seem to provide assessment designers with sufficiently clear guidance on how to design effective application tasks. This is because the descriptor "realistic" can be attached to any task that makes reference to real-world activities or real-world entities, regardless of how strong its links to the real world are. The concept "purposeful", on the other hand, as operationalized in this study, is increasingly more useful as it draws attention to the "why" question. By doing so, it highlights the need for designing tasks that do not only bear resemblance to the real world but are also meaningful. Such tasks can support the application goal of the assessment more effectively.

Toward a more comprehensive framework of contextualization for assessing application of knowledge: combining Interpretations I and II

The four principles that were developed through this research, namely deep contextualization, context balance, context unpredictability and context purposefulness, are *"skill-specific"* in that they relate specifically to the assessment of application of knowledge (see Interpretation II). However, for application of knowledge to be assessed effectively, these four principles need to be employed in conjunction with a series of other principles that relate to examination question writing more generally (see Interpretation I). These latter principles, which will here be referred to as *"skill-agnostic"*, signal practices which, if implemented, can enhance the quality of examination questions irrespective of which skill is assessed. These skill-agnostic principles, which were developed through experimental research (e.g., Crisp & Sweiry, 2006; Fisher-Hoch et al., 1997), corpus analysis (e.g., Beauchamp & Constantinou 2020; Constantinou, 2020, 2023) and interviews with professional examination question writers (e.g., Crisp et al., 2018;

Spalding, 2011), aim to optimize examination questions by enabling them to measure what they are intended to measure. Examples of such skill-agnostic principles include: avoiding complex syntactical structures that may increase the reading demand of questions and therefore disadvantage students who are not native speakers of the target language; formulating questions in an unambiguous manner so that it is clear to students what they are expected to do; avoiding complicated contexts, or contexts that may be unfamiliar to some groups of students; using an appropriate layout and spacing; and ensuring that any

Figure 2. A proposed framework of task contextualization aimed at supporting assessment practitioners in developing and/or evaluating tests that assess application of knowledge



accompanying resources, such as diagrams and graphs, are sufficiently clear. (for an overview of these principles, see Crisp et al., 2018; Crisp & Greatorex, 2023).

The two sets of principles that need to be followed for students' application skills to be assessed effectively, namely *skill-specific principles* and *skill-agnostic principles*, are represented in Figure 3. Figure 3 depicts a framework that combines the findings of this research (i.e., skill-specific principles) with existing knowledge about the construction of high-quality examination questions (i.e., skill-agnostic principles), to provide, or to propose, a more comprehensive approach to assessing application skills through written tests.

It should be noted that the primary aim of this paper is to further illuminate the issue of task contextualization in assessment, and to help identify directions for future research. Its intention is not to make judgements about the quality and effectiveness of the legacy Functional Skills qualification. Rather, it is to present and discuss a set of qualitative observations, reflections and realizations that arose from the attempt to evaluate the degree of alignment between the approach to task contextualization implemented in the Functional Skills examination papers and the claims and objectives of the qualification.

It is hoped that this exploratory study will lead to further research in this area. Such research should seek to refine, extend and validate the proposed theoretical framework by investigating how task contextualization is implemented in other subjects and qualifications around the world. Developing a deeper understanding of task contextualization presupposes identifying – and building a typology of – the different approaches to contextualization currently used in assessment, as well as understanding their functional underpinnings. The present study constitutes a first step in this direction.

References

Anderson, J. R., Reder, L.M., & Simon, H. A. (1996). Situated learning and education. *Educational Researcher, 25*(4), 5-11.

Baird, J., Hopfenbeck, T.N., Elwood, J., Caro, D., & Ahmed, A. (2014). *Predictability in the Irish Leaving Certificate*. Oxford University Centre for Educational Assessment Report OUCEA/14/1.

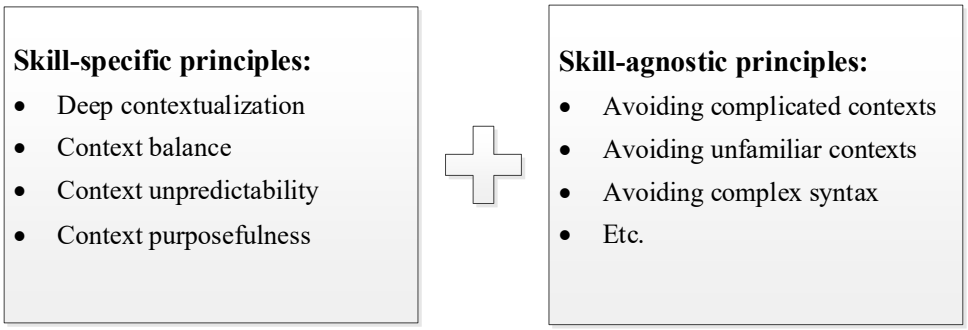
Bakker, A., Wijers, M., Jonker, V, & Akkerman, S. (2011). The use, nature and purposes of measurement in intermediate-level occupations. *ZDM Mathematics Education, 43*, 737-746.

Beauchamp, D., & Constantinou., F. (2020). Using corpus linguistics tools to identify instances of low linguistic accessibility in tests. *Research Matters: A Cambridge Assessment Publication, 29*, 10-16. https://files.eric.ed.gov/fulltext/EJ1293945.pdf

Beswick, K. (2011). Putting context in context: an examination of the evidence for the benefits of 'contextualised' tasks. *International Journal of Science and Mathematics Education, 9*, 367-390.

Bloom, B. S., Engelhart, M., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I. The classification of educational goals - cognitive domain*. Longman.

Figure 3. A more comprehensive framework of contextualization for assessing application of knowledge



Boaler, J. (1993). The role of contexts in the mathematics classroom: do they make mathematics more "real"? *For the Learning of Mathematics, 13*(2), 12-17.

Boaler, J. (1994). When do girls prefer football to fashion? An analysis of female underachievement in relation to 'realistic' mathematic contexts. *British Educational Research Journal, 20*(5), 551-564.

Constantinou, F., Crisp, V., & Johnson, M. (2018). Multiple voices in tests: towards a macro theory of test writing. *Cambridge Journal of Education, 48*(4), 411-426.

Constantinou, F. (2020). Examination questions as a form of communication between the examiner and the examinee: a sociolinguistic perspective on assessment practice. *Cambridge Journal of Education, 50*(6), 711-728.

Constantinou, F. (2023). How novel can examination questions really be? Exploring the boundaries of creativity in examination question writing. *Research Papers in Education, 38*(2), 208-226.

Crisp, V., & Sweiry, E. (2006). Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research, 48*(2), 139-154.

Crisp, V., Johnson, M., & Constantinou, F. (2018). A question of quality: conceptualisations of quality in the context of educational test questions. *Research in Education, 105*(1), 18-41.

Crisp, V., & Greatorex, J. (2023). The appliance of science: exploring the use of context in reformed GCSE science examinations. *Assessment in Education: Principles, Policy and Practice*. Advance online publication. https://doi.org/10.1080/0969594X.2022.2156980

Cumming, J. J., & Maxwell, G. S. (1999). Contextualising authentic assessment. *Assessment in Education: Principles, Policy and Practice, 6*(2), 177-194.

Dearing, R. (1996). *Review of qualifications for 16-19 year olds: quality and rigour in A Level examinations*. School Curriculum and Assessment Authority.

De Lange, J. (1995). Assessment: no change without problems. In T. A. Romberg (Ed.), *Reform in school mathematics and authentic assessment* (pp. 87-172). State University of New York.

Department for Education (DfE). (2014). *The national curriculum in England: Key stages 3 and 4 framework document*. Department for Education. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/381754/SECONDARY_national_curriculum.pdf

Doorman, M. (2019). Contexts to make mathematics accessible and relevant for students – Jan de Lange's contributions to Realistic Mathematics Education. In W. Blum, M. Artigue, M. A. Mariotti, R. Sträßer, M. van den Heuvel-Panhuizen (Eds), *European traditions in didactics of mathematics, ICME-13 Monographs* (pp. 73-78). Springer.

Education and Training Foundation (ETF). (2015). *Making Maths and English work for all: the review of what employers and learners need from the Maths and English qualifications taken by young people and adults*. Retrieved from http://www.et-foundation.co.uk/wp-content/uploads/2015/04/Making-maths-and-English-work-for-all-25_03_2015002.pdf

Fisher-Hoch, H., Hughes, S., & Bramley, T. (1997, September). *What makes GCSE exam questions difficult? Outcomes of manipulating difficulty of GCSE questions*. Paper presented at the British Educational Research Association annual conference, York, UK.

Flick, U. (2018). *An introduction to qualitative research* (6th edn). Sage.

Gainsburg, J. (2005). School mathematics in work and life: what we know and how we can learn more. *Technology in Society, 27*(1), 1-22.

Gravemeijer, K., & Doorman, M. (1999). Context problems in realistic mathematics education: a calculus course as an example. *Educational Studies in Mathematics, 39*, 111-129.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Hastwell, K., Strauss, P., & Kell, C. (2013). 'But pasta is pasta, it is all the same': the language, literacy and numeracy challenges of supermarket work. *Journal of Education and Work, 26*(1), 77-98.

Hodgen, J., & Marks, R. (2013). *The employment equation: why our young people need more maths for today's jobs*. The Sutton Trust. Retrieved from https://www.suttontrust.com/wp-content/uploads/2019/12/MATHSREPORT_FINAL-1.pdf

Hoyles, C., Noss, R., & Pozzi, S. (2001). Proportional reasoning in nursing practice. *Journal for Research in Mathematics Education, 32*(1), 4-27.

Jurdak, M. E. (2006). Contrasting perspectives and performance of high school students on problem solving in real world situated, and school contexts. *Educational Studies in Mathematics, 63*, 283-301.

Kelly, A. (2001). The evolution of key skills: towards a tawney paradigm. *Journal of Vocational Education and Training, 53*(1), 21-36.

Koh, K. (2017). Authentic assessment. *Oxford Research Encyclopedia of Education*. https://doi.org/10.1093/acrefore/9780190264093.013.22

Kramarski, B., Mevarech, Z. R., & Arami, M. (2002). The effects of metacognitive instruction on solving mathematical authentic tasks. *Educational Studies in Mathematics, 49*, 225-250.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: an overview. *Theory into Practice, 41*(4), 212-218.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Long, R. (2017). *Briefing paper: GCSE, AS and A level reform (England)*. House of Commons.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). Macmillan.

Nesher, P., & Teubal, E. (1975). Verbal cues as an interfering factor in verbal problem solving. *Educational Studies in Mathematics, 6*(1), 41-51.

OECD. (2019*). Skills matter: Additional results from the Survey of Adult Skills*. OECD. https://doi.org/10.1787/1f029d8f-en.

OECD. (2023). *PISA 2022 assessment and analytical framework*. OECD. https://doi.org/10.1787/dfe0bf9c-en.

Office of Qualifications and Examinations Regulation (Ofqual). (n.d.). *What qualification levels mean*. https://www.gov.uk/what-different-qualification-levels-mean/list-of-qualification-levels

Office of Qualifications and Examinations Regulation (Ofqual). (2011). *Functional Skills criteria for mathematics: Entry 1, Entry 2, Entry 3, Level 1 and Level 2*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/371154/11-10-07-functional-skills-criteria-for-mathematics.pdf

Office of Qualifications and Examinations Regulation (Ofqual). (2015). *Improving Functional Skills qualifications*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/398441/2015-01-27-improving-functional-skills-qualifications.pdf

Olfos, R., & Zulantay, H. (2007). Reliability and validity of authentic assessment in a web based course. *Educational Technology and Society, 10*(4), 156-173.

Smith, C., & Morgan, C. (2016). Curricular orientations to real-world contexts in mathematics. *The Curriculum Journal, 27*(1), 24-45.

Sokhanvar, Z., Salehi, K., & Sokhanvar, F. (2021). Advantages of authentic assessment for improving the learning experience and employability skills of higher education students: A systematic literature review. *Studies in Educational Evaluation*, *70*, 101030.

Spalding, V. (2011). *Structuring and formatting examination papers: Examiners' views of good practice* (Report). Assessment and Qualifications Alliance. https://filestore.aqa.org.uk/content/research/CERP-RP-VS-05022010_0.pdf?download=1

Steen, L. A. (2003). Data, shapes, symbols: achieving balance in school mathematics. In B. L. Madison & L. A. Steen (Eds), *Quantitative literacy: Why literacy*

matters for schools and colleges (pp. 53-74). The Mathematical Association of America.

Sullivan, P., Zevenbergen, R., & Mousley, J. (2003). The contexts of mathematics tasks and the context of the classroom: are we including all students? *Mathematics Education Research Journal, 15*(2), 107-121.

van den Heuvel-Panhuizen, M., & Drijvers, P. (2014). Realistic mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 521-525). Springer.

van den Heuvel-Panhuizen, M. (2019). Didactics of mathematics in the Netherlands. In W. Blum, M. Artigue, M. A. Mariotti, R. Sträßer, M. van den Heuvel-Panhuizen (Eds), *European traditions in didactics of mathematics, ICME-13 Monographs* (pp. 57-94). Springer.

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: creating a blueprint for course design. *Assessment and Evaluation in Higher Education, 43*(5), 840-854.

Wake, G., & Williams, J. (2003). Using workplace practice to inform curriculum design. In S. Lamon, W. Parker & S. Houston (Eds), *Mathematical modelling a way of life* (pp. 189-200). Horwood Publishing.

Wiggins, G. (1989). A true test: toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703-713.

Wiggins, G., & McTighe, J. (2007). *Schooling by design: mission, action, and achievement*. Association for Supervision and Curriculum Development.

Wijaya, A., van den Heuvel-Panhuizen, M. & Doorman, M. (2015). Opportunity-to-learn context-based tasks provided by mathematics textbooks. *Educational Studies in Mathematics, 89*(1), 41-65.

Wiliam, D. (2008). Quality in assessment. In S. Swaffield (Ed.), *Unlocking Assessment: understanding for reflection and application* (pp. 123-137). Routledge.

World Bank. (2023). *How foundational skills are critical for any occupation (Skills4Dev Knowledge Digest, Issue 5).* Retrieved from https://thedocs.worldbank.org/en/doc/f80b7709ebb5844221fecd2474bb6a1c-0200022023/related/Skills4Dev-June-2023-Foundational-Skills.pdf

Citation:
Constantinou, F. (2024). Assessing students' application skills through contextualized tasks: Toward a more comprehensive framework for embedding test questions in context. *Practical Assessment, Research, & Evaluation*, 29(10). Available online: https://doi.org/10.7275/pare.2103

Corresponding Author:
Filio Constantinou
Cambridge University
Email: filio.constantinou@cambridge.org