Fermi problems in standardized assessment in grade 8

Problemas de Fermi em testes de avaliação padronizados no 8.º ano

Gilbert Greefrath
University of Münster
Germany
greefrath@uni-muenster.de

Lena Frenken
University of Münster
Germany
l.frenken@uni-muenster.de

Abstract. Since 2006, Germany has been pursuing a comprehensive educational monitoring strategy that includes comparative and standardized assessment tests (called VERA) in mathematics. These tests are administered state-wide and, with a few exceptions, in the eighth grade of every general education school. Among other competencies, these tests examine the modelling competency of students. In application and modelling tasks, the various requirements associated with testing tasks create specific challenges that often result in word problems rather than real applications. One possible approach to setting a suitable modelling problem for assessment is to use Fermi problems that draw upon a real context. Based on various classifications of mathematical tasks, this paper develops a series of criteria for Fermi problems for assessment purposes. These criteria are applied specifically to Fermi problems contained in the described standardized assessment tool in Grade 8 in Germany. Based on the findings, differences and similarities between Fermi problems are determined and discussed. Fermi problems exhibit a certain homogeneity as specialized modelling tasks, but are also associated with a broad spectrum of difficulties, which seem to be linked to the number of mathematical quantities required for the solution. Various Fermi problems can cover many different aspects of performance and are a good way to incorporate authentic situations into test problems.
Keywords: Fermi problems; mathematical modelling; task criteria; modelling tasks; test problems.

Resumo. Desde 2006, a Alemanha tem vindo a seguir uma estratégia abrangente de monitorização educacional que inclui testes de avaliação comparativos e padronizados (chamados VERA) em matemática. Estes testes são administrados em todo o país e, com algumas exceções, no oitavo ano

Quadrante: Revista de Investigação em Educação Matemática 30(1) 52-73
https://doi.org/10.48489/quadrante.23587
Received: November 2020 / Accepted: May 2021

de todas as escolas de ensino básico. Entre outras competências, estes testes examinam a competência de modelação dos estudantes. Nas tarefas de aplicação e modelação, os vários requisitos associados à construção de tarefas para testes de avaliação criam desafios específicos que muitas vezes levam a que tais tarefas resultem em problemas de palavras em vez de aplicações reais. Uma abordagem possível para criar um problema de modelação adequado para um teste de avaliação é utilizar problemas de Fermi, que se baseiam num contexto real. Tendo por base várias classificações de tarefas matemáticas, este artigo desenvolve uma série de critérios para criar problemas de Fermi para fins de avaliação. Estes critérios são aplicados especificamente aos problemas de Fermi incluídos no referido instrumento de avaliação padronizada, no 8.º ano, na Alemanha. Com base nos resultados, as diferenças e semelhanças entre diversos problemas de Fermi são identificadas e discutidas. Os problemas de Fermi apresentam uma certa homogeneidade como tarefas específicas de modelação, mas estão também associados a um amplo espetro de dificuldades, que parecem estar ligadas ao número de quantidades matemáticas necessárias para obter a solução. Diferentes problemas de Fermi podem abarcar muitos aspetos diferentes de desempenho e são uma boa forma de incorporar situações autênticas em problemas de teste.
Palavras-chave: problemas de Fermi; modelação matemática; critérios de tarefas; tarefas de modelação; problemas de teste.

Introduction

In Germany, in response to the unsatisfactory findings of the PISA study in 2000, the Conference of Ministers of Education decided in 2003 to introduce new educational standards for mathematics and other subjects. One objective of these standards is to offer guidance to teachers, parents, students, and educational administrators. Another objective is to provide a basis for evaluation at various levels, such as the educational system as a whole, schools, or classes. In mathematics, the specific chosen tasks play an essential role in this assessment context, since the competencies described in the national educational standards are concretized by tasks that require these competencies. Although, for test purposes, each task must be individually solvable and the tasks must be reliably codable, the recommendations are not limited to closed task formats. The aim is to propose a wide range of competency-oriented tasks in both classroom and assessment settings (Blum, Drüke-Noe, Leiß, Wiegand, & Jordan, 2005).

In Germany, application-oriented problems are part of everyday mathematics lessons. Mathematical modelling is included as a competency in the national educational standards and hence in the curricula of all federal states. There are many research projects on the topic of mathematical modelling, with a significant increase over the past decade (Greefrath, 2020). However, designing test problems with a real-world context can be particularly challenging if – compared to word problems – the aim is to create open and authentic tasks that facilitate diagnosis. The impact of items utilized for standardized assessment on the tasks deployed in lessons also needs to be taken into consideration. In any case, extra-mathematical contexts and questions about them need to be consistent, generally credible, and familiar to learners, even in the context of assessment. This is essential in order for learners to take these contexts, and the use of mathematics in them, seriously and consider them meaningful. One approach to developing this type of test problem is to use Fermi problems. This article contains a detailed analysis of Fermi problems from the standardized assessment conducted in German eighth grade. The results may be relevant for task design that aims at both creating comparable, standardized items and assessing students' mathematical competencies or thinking. We start with a description of the German situation in the area of standardized assessment in mathematics and then present a characterization of Fermi problems. On the basis of issues regarding real-world tasks within standardized assessment, we develop criteria that can be used as a theoretical classification scheme for such tasks.

Standardized assessment in mathematics

Since 2006, the Conference of Ministers of Education has been pursuing a comprehensive educational monitoring strategy in Germany, whose goal is to reinforce the focus on competencies within the educational system. The participation of Germany in international school performance studies (PISA, since 2000) was in 2006 announced as one part of this monitoring strategy. A central review is also conducted by the German Institute for Educational Quality Improvement (IQB) to determine whether the national educational standards have been achieved. To this end, state-level comparisons are performed at grades 3 and 8 at regular intervals, to gain information about the extent to which students have achieved the competency expectations outlined in the national education standards. The state-level comparison in mathematics and the PISA studies should not be regarded as interchangeable with equivalent scope. Nevertheless, their general concepts and specific tasks have many similarities. There is even a relationship between the competency measurements found by both studies, and very similar test result distributions can be observed at the level of the entire population (Ehmke, Köller, & Stanat, 2017). In addition to the above measures, so-called *comparative tests* (VERA) are also conducted in mathematics among other. These tests are held state-wide in Grade 8 (and other grades) of many general education schools, to investigate the competencies developed by students. The test items used in these comparative tests are developed by teachers from all federal states under the direction of the German Institute for Educational Quality Improvement. In a multistep revision process, the tasks are reviewed and optimized after the evaluation by maths education experts from universities. Before being deployed in a comparative test, the tasks are tested in a pilot of several hundred students. Tasks are only used in an actual comparative test if they are found to be suitable (see www.iqb.hu-berlin.de/vera). Unlike the national and international school performance studies, the comparative tests are graded by teachers before reporting the results to state-level institutions. Feedback about the results of the comparative tests is carefully prepared, intended to be used by teachers for lesson development, as well as to provide diagnostic information for learners and their parents. Comparative tests can influence subsequent lesson planning by teachers, primarily through the choice of task formats. The acceptance of these comparative tests among teachers varies, and can certainly be improved, for example by allowing teachers to participate in the processes of selecting test contents and be trained in the use of test results for instructional development (Maier, Metz, Bohl, Kleinknecht, & Schymala, 2012).

Task design for assessment

As mentioned in the previous section, the tasks for the VERA are developed very carefully and in compliance with various criteria. Especially the aims of task usage should be considered. Some tasks, created to support the learning process, have a more open format. These tasks aim at learning skills and developing competencies, and can therefore be described as learning tasks. In most cases, they do not provide all the information needed to find a solution, and students must first select and research their information independently. If a task is created for assessment, it should be designed to enable scoring as clearly, comparably, and simply as possible. One of the aims of such an assessment task is to determine the performance and state of competency development of learners accurately as well as reliably. This leads to psychometrically defined prerequisites of test items. For example, the key objective of an item has to be clear and unambiguous. In addition, the emergence of measurement errors should be minimized (van den Heuvel-Panhuizen & Becker, 2003). Furthermore, items or tasks should be formulated as clearly as possible so that the evaluation process can yield consistent information and valid interpretations of results (Joint Committee on Standards for Educational Evaluation, 2018). Therefore, the item construction process for the German comparative tests spans a wide range and involves various experts (Schecker & Parchmann, 2007). The abovementioned properties often lead to the assumption that standardized tasks should only have one correct solution or that the data needed to solve the task should be provided (van den Heuvel-Panhuizen & Becker, 2003). However, it is not always possible or even desirable to distinguish strictly between learning and testing tasks (Blum et al., 2005; Büchter & Leuders, 2005). It should also be borne in mind that assessment tasks could have an influence on which teaching tasks are selected for instructional purposes (Prodromou, 1995).

Though it is generally possible to distinguish between the goals of assessment and learning, the corresponding tasks cannot be clearly separated; some assessment tasks can be used for both diagnosis and performance. However, assessment tasks aim at collecting as much information as possible in a given, usually, short-time period. Furthermore, the distinction between learning tasks and testing tasks does not yield any clear insight into the cognitive characteristics of the tasks. For example, a study of assessment tasks designed at the class level in Germany showed that modelling only played a minor role in solving real-world tasks, whereas technical working was consistently important (Drüke-Noe, 2014).

During the design process of items for the comparative tests various obstacles, barriers, and challenges occur: it is difficult to train the item writers; scoring of certain item types is associated with substantial amounts of time and cost; and the range of competencies that are assessed within one test is limited (Schecker & Parchmann, 2007). In particular, designing assessment tasks with a real context and real-world applications can be difficult, if the goal is to create an open, authentic task that facilitates diagnosis rather than merely a word problem. The diverse requirements associated with standardized assessment create specific challenges for tasks involving application situations and modelling, often leading to technical exercises or word problems rather than real applications in test settings. One possibility for coping with the described issue on task design within standardized assessment might be to use Fermi probems based on real contexts in tests.

Fermi problems

Fermi problems are open estimation problems with real-world contexts that can be used with the pedagogical purpose of clearly identifying assumptions and encouraging students to estimate certain quantities or variables. The required information can typically be processed without pen and paper (Bergman Ärlebäck, 2009; Sriraman & Lesh, 2006). With regard to the task design, additional properties can also be included in Fermi problems in mathematics lessons. For example, they can be developed as motivating, challenging, and open tasks that do not contain any numbers, so as to avoid students performing calculations without thinking about the context (Peter-Koop, 2004). Fermi problems can also typically be broken down into smaller problems that can be solved by estimating (Albarracín & Gorgorió, 2014; Ross & Ross, 1986). Fermi problems can therefore be characterized as a special case of modelling tasks. Thus, unlike word problems, Fermi problems require the processing of several steps from the modelling cycle. In contrast and in line with Niss and Blum (2020), we understand word problems as questions posed "in a real, idealised or imagined extra-mathematical context and situation, the answering of which requires some sort and degree of mathematical problem solving" (Niss & Blum, 2020, p. 30). In contrast to modelling and especially to Fermi problems, word problems do not require pre-mathematisation, de-mathematisation or validations of results and the entire model. From the perspective of problem solving, Fermi problems are usually underdetermined, open tasks with a final state that is clearly formulated in a question but an unclear initial state and an unclear transformation (e. g. Pólya, 1981). They focus strongly on data acquisition through multiple estimates. Fermi problems are named after the nuclear physicist and

Nobel Prize winner Enrico Fermi (1901-1954), who was known for performing quick estimates of problems with virtually no available data. The classic example of a Fermi task is the question of how many piano tuners there are in Chicago (Peter-Koop, 2004). At first, there seems to be no information available to answer this question. However, it is possible to gradually estimate the order of magnitude to be around 100, by making reasonable assumptions about the number of residents in Chicago, the size of a household, the proportion of households with a piano, the period between two piano tunings, the duration of a piano tuning session, and the workload of a piano tuner, thereby constructing a meaningful answer to the question. The answer is found by making suitable selections and reasonable estimates about intermediate information. Niss and Blum (2020) summarize: "While they [Fermi problems] are, in principle, closed when it comes to the kind of answers sought (mostly just a number or a quantity), they are extremely open when it comes to choosing a framework within which they can be specified and dealt with". Thus, Fermi problems are characterized by their particularly high accessibility, and they can be implemented at various educational levels (Ferrando & Albarracín, 2021; Peter-Koop, 2009). Above and beyond inspiring further questions, they can motivate the use of mathematics in the real world. Fermi problems are regarded as very useful in mathematics, especially for promoting problem-solving and modelling (Bergman Ärlebäck & Albarracín, 2019).

When using Fermi problems in mathematics lessons, there is less emphasis on inner-mathematical aspects than on other steps of the modelling process, such as mathematizing and validating. In particular, Fermi problems can be used to address managing inaccuracy caused by estimation, which is often not granted much attention in mathematics lessons (Bergman Ärlebäck & Bergsten, 2010).

The concept of a Fermi task is also used in a broader sense to denote open tasks containing only a question or a few pieces of information, such as an image. For our purposes, problems that can be solved without an illustration by estimating the intermediate data, like the question of how many piano tuners there are in Chicago, are referred to as original Fermi problems. Thus, original Fermi problems place particular emphasis on estimating and working with imprecise information. The translation into mathematical models that are as simple as possible also play an important role. By contrast, Fermi problems in the broader sense can also encompass research and experimentation, as well as finding different ways to approach a problem. Students also learn to ask questions themselves and work with heuristic strategies. They learn to apply their everyday knowledge in order to perform calculations involving quantities (Greefrath, 2019b). In this paper, Fermi problems are understood in the broader sense. In essence, a definition of Fermi problems could be the following: Fermi problems consist of an open question or a few pieces of information, and possibly a picture from the real world, whereby the question targets using knowledge of the world, estimations and measurements, so that the solution path offers different possibilities, but the final result usually only contains a number or quantity that has to be interpreted in the given context.

Classification of Fermi problems in assessment

In this section, we wish to deduce possible aspects for a classification scheme of Fermi problems that can be posed within standardized assessment. Therefore, we consider aspects of task design for assessment and the abovementioned challenges on the one hand and properties of Fermi problems and reality-based problems on the other hand.

There are many ways to classify mathematical tasks. For example, the OECD assessment framework (2013) defines mathematical topics (e.g. geometry), required skills (e.g. defining, solving, reasoning), task complexity (e.g. one-step or multi-step), and task format (e.g. multiple choice). In follow-ups to the PISA 2000 study, various other difficulty-influencing characteristics of test items were also identified. Examples include modelling complexity, curricular knowledge level, contexts, and openness (Neubrand, Klieme, Lüdtke, & Neubrand, 2002), as well as basic mental models (Blum, vom Hofe, Jordan, & Kleine, 2004). A comprehensive classification of tasks specifically tailored to modelling was developed by Maaß (2010). This classification encompasses a wide range of modelling tasks and provides a foundation on which we can classify Fermi problems. In this study we focus on Fermi problems that firstly, are encountered in the context of standardized assessment, and secondly, can be characterized as specialized modelling tasks. Accordingly, we consider classifications of both assessment and modelling tasks. The theoretically-based characteristics of Fermi problems in assessment are then discussed from four different perspectives.

Dimensions in educational standards

Assessment tasks for the secondary level can encompass a wide range of mathematical content. Consequently, it makes sense to organize tasks according to the subject areas studied at the secondary level (e.g. stochastics, arithmetic, algebra, geometry). Some tasks can be classified into multiple subject areas. Similarly, we can use the classification into key ideas [Leitideen] introduced in the German educational standards (KMK, 2012; Schecker & Parchmann, 2007). The five key ideas are numbers, measuring, space and shape, functional relationships, data and randomness. These key ideas are intended to emphasize fundamental concepts and demonstrate that mathematical phenomena such as counting or measuring inspire the development of mathematical subjects. The two criteria of subject areas and key ideas allow us to structure tasks according to the specialized mathematical content that they contain.

In parallel, it is useful to look at the potential solving processes associated with the various tasks. Although solving a task depends on a variety of factors and not just the task itself, we can focus on the solving process anticipated when the task was designed. Here, we can consider general competencies as a criterion for structuring mathematical processes. Besides the five mathematical processes of arguing mathematically, solving problems mathematically, using mathematical representations, working technically, communicating mathematically, the competency of mathematical modelling also belongs to this category (for explanations on the mentioned competencies see Niss & Højgaard, 2019).

General task criteria

The *openness* of a task is another general criterion whose usefulness is not limited to tasks with a connection to reality. An open task can be defined as offering several solutions, either by not setting a clear initial or final state, or by enabling learners to use different solution paths or transformations of the initial state (Greefrath, 2010). A classification into clear and ambiguous tasks only provides a very rough description of the situation. These concepts contain both a subjective component and an objective one. The subjective component implies some dependency on the ability of the learner. The objective component relates to whether specific information is missing in the task and is therefore potentially only accessible with limited precision. Fermi problems typically have open initial states and possibly also an open transformation.

In addition to these general criteria, there are also specific criteria that are only applicable to tasks with a real context, such as Fermi problems. The text of a task can contain *information* that is not required in order to solve the task. If so, the task is defined as overdetermined. The reverse case is also possible, in which case the task does not contain all information required for solving. Such a task is then underdetermined. The missing information needs to be determined, for example by drawing upon everyday knowledge, performing an estimate, or conducting research. A task could also conceivably be both types at once. For example, if some unnecessary information is specified in the task, but other necessary information is not provided and needs to be researched, estimated, or otherwise obtained, the task is both underdetermined and overdetermined. This criterion is considered relevant from both points of view, the task as an element of a standardized assessment instrument and the task with a real context.

Contextual criteria

In addition to the characterization described above, the connection to reality of a task can be defined more precisely using concepts such as *authenticity*. An authentic task is both plausible for students and realistic with respect to its environment. Authenticity implies that of both the extra-mathematical context and the way that mathematics is applied in a particular situation. The extra-mathematical context must be realistic and not specifically invented or contrived for the mathematics problem. The way that mathematics is used in the problem situation must also be sensible and realistic, and not contrived solely for a mathematics lesson. In authentic tasks, students can assume that they are working with things that actually exist, and that the task or problem being formulated is therefore a real problem (Vos, 2018) which is justified beyond mathematics lessons. Authentic modelling tasks entail problems that could be encountered in a field or problem area of the real world and would be accepted by someone who works in this area (Niss, 1992). Authenticity helps students to take a task seriously and therefore avoid superficial replacement strategies for solving them, as can occur with word problems (Palm, 2007).

Task authenticity does not necessarily imply that the learners must have an actual need for the corresponding applications, or that these tasks are or will be important in their current or future life. We can therefore also consider the *relevance* of a task for learners. Considering students as participating socially in an environment, a task is relevant if it is important for the learner's current or future life (Hernandez-Martinez & Vos, 2018). The property of a task that is already considered important from the learner's perspective can be described as student relevance. On the other hand, if a task will only become relevant for students in the future, it can be said to have life relevance (Greefrath, Siller & Ludwig, 2017).

In modelling cycles such as those described by Blum & Leiss (2007), sub-competencies are also taken into account in modelling activities. The ability to execute a certain process can be viewed as a sub-competency of modelling (Kaiser, 2007; Maaß, 2006). Modelling processes can be divided into the sub-competencies of understanding, simplifying, mathematizing, working mathematically, interpreting, validating, and communicating. If another modelling cycle were used as the basis, a set of sub-competencies with different emphases could conceivably be used.

Criteria on the involved magnitudes

In her comprehensive classification scheme, Maaß (2010) considers many of the criteria presented above, including the type of connection to reality, the openness of a task, and the focus on modelling activities. In the specific context of Fermi problems, we are also interested in which and how many *magnitudes* have to be determined by the test takers in order to solve the realistic problem. Values for these magnitudes are typically determined by estimating, measuring, calculating, or extracting information from the text (reading). By making assumptions on typical students' solutions, those magnitudes can be counted and analysed.

With regard to the *representation* of a magnitude being estimated in a task, a given quantity can be presented as an object, as an image, or merely abstractly as a description in the text. The way in which the estimated magnitudes are presented is potentially source of difficulty. As a general rule, representations as objects or photos tend to be more accessible than asking students to estimate magnitudes that are only presented abstractly (Maaß, 2010).

When estimating a magnitude, a certain level of reference knowledge about familiar magnitudes or knowledge of the world might be needed for mental comparisons, such as the height of a typical adult. These known magnitudes are *reference values* for students. The type of quantity is also relevant here; for example, estimating numbers and lengths is presumably more familiar to students than estimating weights. The number of magnitudes that needs to be estimated is also a characteristic property of a task.

Based on the considerations presented above, a selection of criteria for Fermi problems within standardized assessment is summarized below (Table 1). After formulating the research question, we present three example tasks and use this table to classify one of them according to all criteria.

Table 1. Criteria for Fermi problems in standardized assessment

Criterion	Characteristic questions
Subject area	Which subject area does the task emphasize? (stochastics, arithmetic, algebra, geometry)
Key idea	Which key idea does the task primarily emphasize? (numbers, measuring, space and shape, functional relationships, data and randomness)
General competency	Which general competency is needed to complete the task? (arguing mathematically, solving problems mathematically, modelling mathematically, using mathematical representations, working technically, communicating mathematically)
Openness	Does the task have an ambiguous initial or target state / an ambiguous transformation?
Information	Is the task overdetermined, underdetermined, or both?
Authenticity	Does the problem relate authentically to the physical world / the use of mathematics?
Relevance	Is the problem currently relevant to students (student relevance) / will it be relevant later in life (life relevance)?
Modelling sub-competencies	Which modelling sub-competencies are primarily targeted when working on the tasks? (understanding, simplifying, mathematizing, working mathematically, interpreting, validating, communicating)
Magnitudes	Which and how many magnitudes need to be determined by estimating, measuring, reading, calculating?
Reference concepts	What reference concepts about quantities are required? (e.g. size of a person, height of a door, volume of a milk carton, etc.)
Representation	Is the quantity being estimated represented by an object, a picture, or just text?

Research questions and methodology

The goals of comparative tests (VERA) in Germany are on the one hand to investigate which competencies have been acquired by students at a certain point in their education, for example Grade 8 (aged 13-14), and on the other hand to give teachers differentiated feedback on the requirements faced by their students with respect to national educational standards.

Fermi problems have regularly been used in German comparative tests in Grade 8 (VERA 8), alongside other types of task. On the basis of the theoretically considered criteria, we now focus on whether there is one particularly suitable way to characterize Fermi problems in assessment settings. This would provide valuable information to anyone charged with task design. We therefore pose the following research question: Which criteria do the Fermi problems used in comparative tests in Grade 8 satisfy?

To answer this question, the nine Fermi problems in comparative tests in Grade 8 from 2015 to 2018 were examined in terms of the criteria presented above. The tasks were examined and evaluated individually, using a structural content analysis based on the deductively established criteria presented in the classification section by two independent raters (Mayring, 2014). With regard to more clearly identifying the quantities involved in each task, another analysis was performed for each Fermi task by extracting the underlying structure of typical standard solutions (Reit & Ludwig, 2015b, 2015a).

An overview of the tasks is presented in Table 2. A detailed presentation of three example tasks is given below. Since the original tasks may not be printed, we describe the tasks and insert a comparable image there.

Table 2. Overview of the nine analysed Fermi problems and their difficulty level

Task context	Question	Difficulty level[1]
Bee	How long is the depicted bee?	0,02
Chocolate smarties	Estimate how many chocolate smarties are shown in this photo.	0,36
Rhino	How long is the depicted rhino?	0,48
School building	How tall is the school building?	0,50
Sculpture of a head	Approximately how tall would a giant be to which this part of the head belongs?	0,66
Perfume consumption	How many millilitres of perfume are used up after ten days if a pump is used daily? How long does a 100ml bottle last if several pumps of perfume are used daily.	0,67
Locks	Approximately how many locks are in the photo?	0,67
Distance to school	Is it possible that the total length of the daily routes to school of all students of one German school is as long as the given distance from earth to moon?	0,78
Dustbin	Determine the approximate volume of the dustbin shown.	0,87

Example task – *Chocolate Smarties*

One of the tasks included in a standardized test in 2016 is based on an estimation problem from the museum *Mathematikum* in the German city of Gießen. In the museum, visitors are asked to estimate the number of chocolate smarties shown in the picture with the help of various shapes (triangle, circle, rectangle, square) (Figure 1).

Figure 1. Example task – *Chocolate Smarties* (*Mathematikum Gießen*)

This was converted into a task for standardized assessment by drawing these four shapes onto an image of chocolate smarties. The task was formulated as follows: "Estimate how many chocolate smarties are shown in this photo. You can use any of the shapes to help you". As well as stating their estimate of the number of chocolate smarties, the students were asked to describe their approach. For the evaluation, an interval of correct answers was specified for the student's solutions, and the students were expected to give a description of their solution. The analysis of this task can be retrieved from Table 3.

Table 3. Analysis of the task *Chocolate Smarties*

Criterion	Specification for the example task *Chocolate Smarties*
Subject area	Mainly arithmetic, since the number of chocolate smarties needs to be determined
Key idea	Measuring, since a shape in the photo needs to be used
General competency	Mathematical modelling, because the real situation needs to be translated into a mathematical model before working within this model. Finally, the results have to be translated back into the real setting.
Openness	The shape can be chosen so that multiple solutions are possible and this task is partly ambiguous. It can be deduced that the problem has an open initial state and an ambiguous transformation.
Information	The task is neither underdetermined nor overdetermined, since all necessary information can be found in the illustration?
Authenticity	The context of the problem is authentic, but it does not seem plausible that experts would use this mathematical method to solve the problem.
Relevance	For students, determining the number of chocolate smarties in this way is not a problem that might be relevant to them in their current or future life.
Modelling sub-competencies	Since a full modelling cycle must be completed to solve the task, the task targets the sub-competencies of simplifying, mathematizing, working mathematically, and interpreting.
Magnitudes	When using the triangle or a rectangle to count the chocolate smarties, two quantities are determined, namely the number of chocolate smarties within the shape and a factor corresponding to how many copies of this shape are needed to cover the entire picture (see Fig. 2).
Reference concepts	/
Representation	Picture

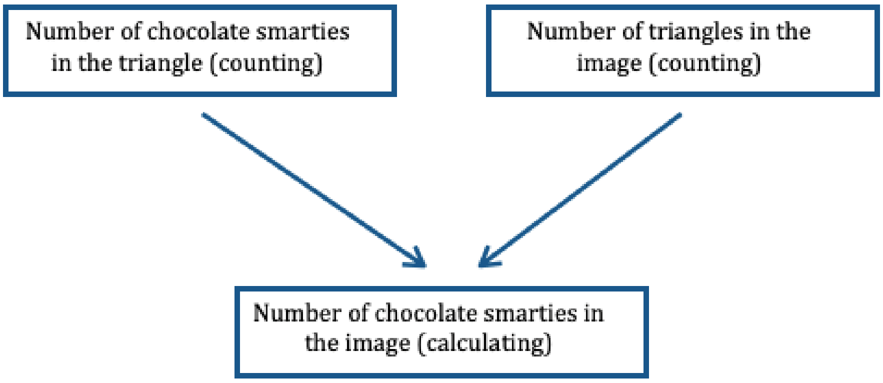

Figure 2. Structure of the *Chocolate Smarties* task

Example task – *School Building*

In the *School Building* task from 2018, a student wishes to find out the full height of his school building after measuring the height of a window. The front of the school is shown in a photo (Figure 3).

Figure 3. Example task – *School Building*

Within the assessment, the following task with a similar photo was posed: "Sasha wants to find out roughly how tall his school is (without the roof). In his classroom, he measured a window and found that it is 2.20 m high". As well as giving the height of the school, test takers are asked to describe their approach. This task can mainly be classified under the subject area of arithmetic and the key idea of measuring, since the solution requires the height of the school to be measured on the photo with a ruler to determine a factor with which to multiply the height of the window. Solving this task requires the general competency of mathematical modelling, since the whole modelling process has to be completed.

To solve the task, the height of the building and that of the window must first be measured in the picture to determine the corresponding factor. Alternatively, it could be directly estimated how often the height of the window fits into the height of the building. For both cases it is possible to derive that the initial state of the task is partly ambiguous. This is also supported by the fact that the problem can be considered to have an open initial state. The task is neither underdetermined nor overdetermined, since all necessary information is in the illustration. The context of the problem is authentic, but it does not seem plausible that experts would use this mathematical method to solve the problem. For students, although a task about a school building appeals to a familiar context, finding the height of an unfamiliar school building is not a problem that is likely to be relevant to them in their current or future life. Solving this problem requires modelling sub-competencies, as a complete modelling cycle must be performed. When measuring the height of the building and that of the window, two magnitudes are determined. The specified height of the window must be extracted from the text of the task. The height of the school building must then be calculated from these three quantities (Figure 4).

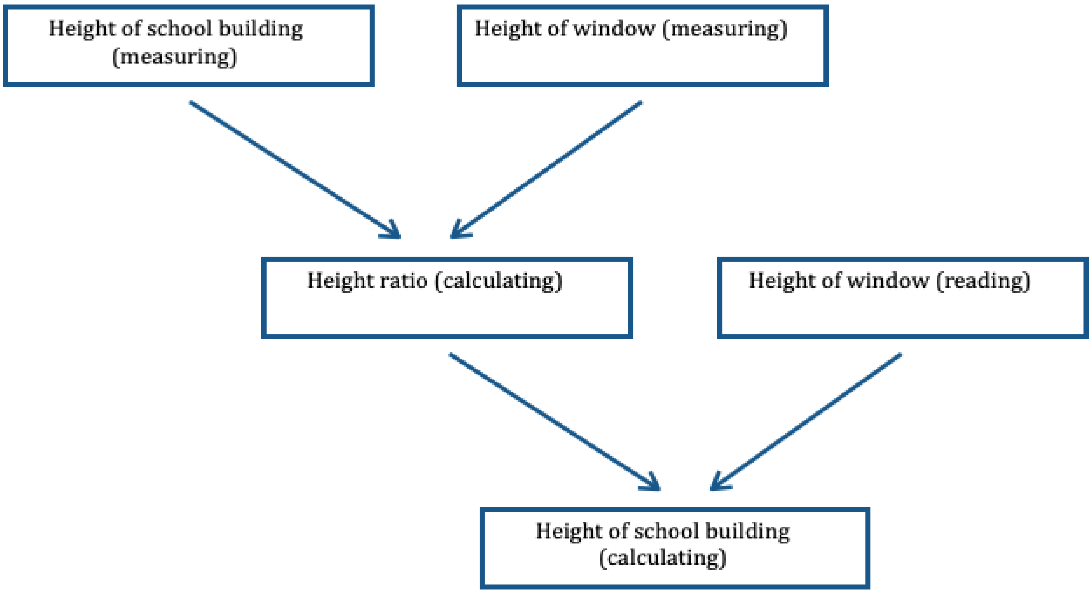

Figure 4. Structure of the *School Building* task

Example task – *Dustbin*

The *Dustbin* task from 2017 is another example of a Fermi task within a standardized assessment instrument. The idea behind this task is to ask students to estimate the volume of a dustbin shown on an image. The task includes a photo showing a woman standing next to a dustbin which has an approximately square base. The task was formulated as follows in the comparative tests: "Determine the volume of this dustbin". As well as stating the volume, students are asked to describe how they solved the task. For the evaluation, an interval of correct numerical answers was specified, and the description of the used representative were coded since these aspects imply a correct answer for the volume.

This task can be assigned to the subject areas of arithmetic and geometry and to the key idea of measuring, since the length, width, and height of the dustbin in the photo need to be estimated using an idea of the person's height to solve the task. The height of the person is estimated first, and the height of the dustbin is then deduced from this estimate. It can be assumed that the width and length of the dustbin are the same. This also allows the length and width of the dustbin to be estimated. The volume of the dustbin can then be calculated from this information.

Solving this task requires the general competency of mathematical modelling, since the whole modelling process has to be completed. To solve the problem, the height, length and width of the dustbin must first be determined on the photo. This is partly ambiguous. There are also several possible ways to solve the problem. For example, it could alternatively be solved directly by comparing the volume with another quantity. Therefore, the problem can be said to have an open initial state and an ambiguous transformation. The task is underdetermined, but not overdetermined, since not all necessary information is given in the text or the illustration, but no additional information can be found. The context of the problem is authentic, but it does not seem plausible that experts would use this mathematical method to solve the problem. For students, determining the volume of a dustbin in this way is not a problem that might be relevant to them in their current or future life. Modelling sub-competencies are required to solve the problem. Since a full modelling cycle must be completed to solve the task, the task targets the sub-competencies of simplifying, mathematizing, working mathematically, and interpreting. The number of magnitudes to estimate depends on the approach. The solution path of estimating height, length and width requires three magnitudes to be determined.

Results

The evaluation of the nine Fermi problems within the standardized assessment instrument for evaluating the national educational standards in Germany shows that all items can be assigned to the subject area of arithmetic, primarily associated with the key idea of measuring. As expected, in terms of general skills, mathematical modelling was identified as the most important process in almost every case. But other skills such as communicating or calculating were also important in some of the tasks. Any openness of the task usually consisted of an unclear initial situation and sometimes of multiple solution possibilities. The unclear initial situation mostly took the form of information missing from the text of or the picture belonging to the task. Accordingly, the information status of many of the tasks was – at least somewhat – underdetermined. But there were also two tasks that cannot be described as either underdetermined or overdetermined. Authenticity was achieved in almost every task through the presence of an authentic situation. Two tasks featured instead, a somewhat hypothetical problem; for example, the task asking whether the distance travelled to school by all of its students can be longer than the distance from the Earth to the Moon, which cannot be said to have an authentic context.

The application of mathematics cannot be considered authentic in almost all the cases. None of the tasks were currently relevant to learners. Two tasks could be described as having life relevance, namely the two relating to spraying perfume and the volume of a dustbin. Completing a full modelling cycle was necessary to solve most of the tasks. Consequently, all sub-competencies of modelling were present in almost every solving process. One task focused only on estimating the length of a bee using a photo. For this task, a complete modelling cycle does not seem to be required; simplification is sufficient to obtain a successful estimate. Since we can conclude that the criteria relating to magnitudes vary once we present more detailed results on the number and type of magnitudes that must be determined to solve the problem in Table 4. Furthermore, Figure 5 shows the number of magnitudes to be determined (ranked by level of difficulty).

Table 4. Fermi problems in comparative tests

Context	**Length of a bee**	**Number of chocolate smarties**	**Length of a rhino**	**Height of a school building**	**Height of a sculpture**	**Amount of perfume consump-tion**	**Number of locks**	**Distance to school**	**Volume of a dustbin**
Quantities Reference concepts	1 length of bee				3 length of body, distance from nose to top of head, height of shape or tree	1 number of perfume sprays per day		2 max. distance to school, number of school weeks	1 height of a person
Quantities Counting		2 number of chocolate smarties per shape, number of shapes					3 number of locks horizontally, number of locks vertically, number of missing locks	1 school day per week	
Quantities Measuring			2 length, height of rhino	2 height of windows, height of building					3 height of person, height of dustbin, width of dustbin
Quantities Reading			1 height of rhino	1 height of window		2 volume of spray, volume of perfume		1 distance from Earth to Moon	
Quantities Calculating	0	1	2	2	2	2	2	3	6

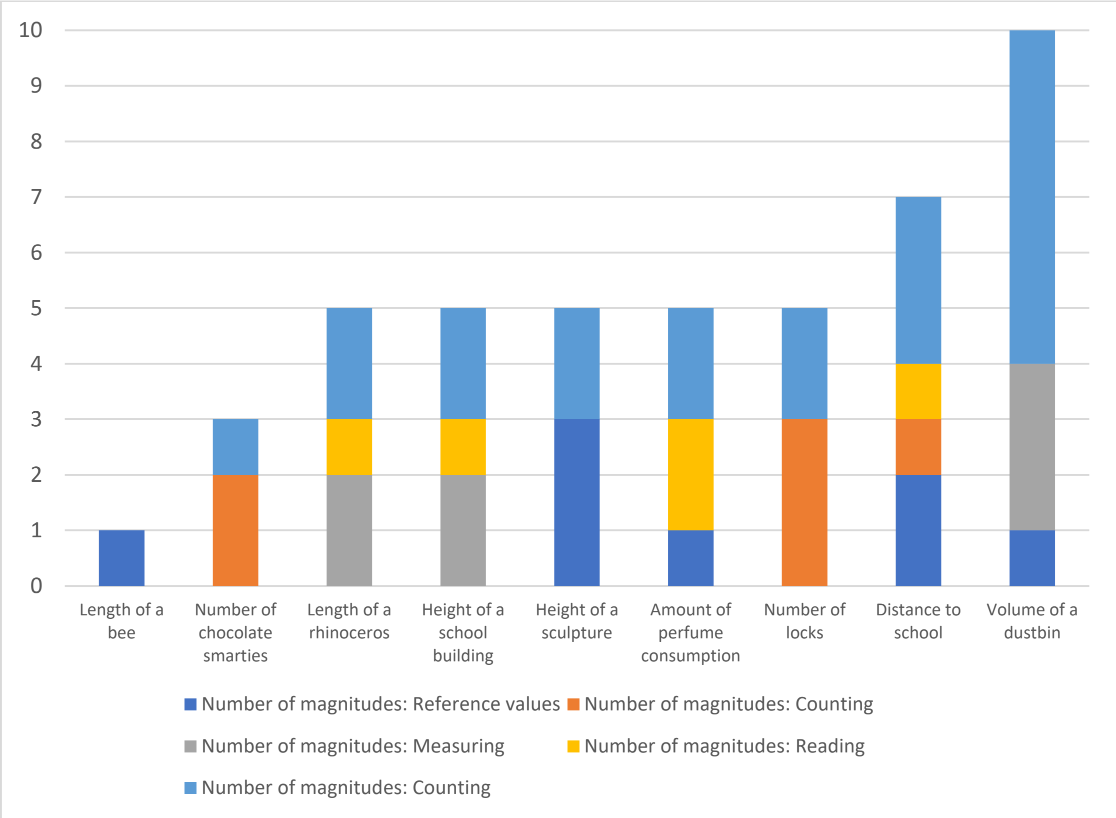

Figure 5. Number of magnitudes to be determined in each task

Discussion

The Fermi problems appear to be quite homogeneous in terms of the subject areas, key ideas, and general competencies that they target. The openness and information of the tasks are also similar, and these two factors depend on one another. Authenticity and the need to complete a full modelling cycle to solve the problem are also frequent. One aspect that could possibly be criticized is the lack of relevance for students. None of the tasks were currently relevant, and only two would be relevant to learners later in life. But overall, this task homogeneity may also be because they were selected as Fermi problems. Many properties of Fermi problems from the literature can be recognized in them (Bergman Ärlebäck, 2009; Peter-Koop, 2004). In particular, the task structure analysis (Reit & Ludwig, 2015a) has shown that Fermi problems can be broken down into smaller problems (Albarracín & Gorgorió, 2014), and that estimating plays an important role (Bergman Ärlebäck & Bergsten, 2010). The properties outlined above also allow us to confirm the definition of Fermi problems as *underdetermined modelling tasks from the subject area of arithmetic that have unclear initial states and contain authentic situations*, especially since the current imprecise definition of Fermi problems has been criticized (Albarracín & Gorgorió, 2014).

In contrast, even though most criteria of Fermi problems are comparable, the range of magnitudes and the methods used to determine solutions vary through the analysed tasks. The hypothesis that the difficulty depends on the magnitudes that must be determined by counting, estimating, measuring, reading, or calculating may be proposed. It appears that the task difficulty can be increased by asking learners to determine a larger number of quantities. As can be seen from the figures on the task structure, the number of combinations that have to be conducted with the determined quantities may also influence the item difficulty.

The homogeneity of the tasks described above is presumably a consequence of their nature as assessment tasks. This can be expected to have prompted a certain standardization in terms of solving time, task format, and compliance with the requirements of the national educational standards. However, it should be noted that the openness and ambiguity are quite unusual for such assessment purposes – especially when bearing in mind most of the associated psychometrical prerequisites (van den Heuvel-Panhuizen & Becker, 2003).

Due to the selection of Fermi problems, it is plausible that the analysed tasks display similar properties and characteristics. The openness, relation to reality, and authenticity were similar across all tasks.

The analysed Fermi problems were mostly represented with a really short text and an additional picture. This contributes to important assessment properties, such as unambiguous and linguistically simple tasks, so that result interpretations can be interpreted clearly and linked to mathematical competency (Joint Committee on Standards for Educational Evaluation, 2018).

Other criteria – such as those that are not typical of modelling or Fermi problems – could possibly play an important role in task design and generating a variety of item difficulties. For instance, basic mental models (Blum et al., 2004) and curricular knowledge level (Neubrand et al., 2002) evidently constitute general factors that influence the difficulty of assessment tasks. Maier et al. (2010) also developed a general didactic category system for analysing the cognitive potential of tasks. Categories such as openness and forms of representation are included, among other aspects. These factors are taken into account by the criteria considered here. However, it was also found that verbal complexity in particular can play an important role in the difficulty of a task. Given this background and the present results about tasks, these criteria for Fermi problems describe possible links to difficulty level but cannot be expected to provide a complete explanation that fully predicts the task difficulty. On a positive note, this means that we have the opportunity to construct truly authentic and relevant tasks for tests with varying degrees of difficulty. This could help to reduce tasks involving technical work, which remain very common in Germany (Drüke-Noe, 2014), and thus improve the rather low acceptance of comparative tests (Maier et al., 2012). However, since comparative tests cover the entire range of national educational standards for secondary school, Fermi problems can only account for a small portion of them. Nevertheless, based on their positive characteristics, the use of Fermi problems in comparative tests and therefore also for lesson development, seems highly desirable.

Conclusion

A detailed study of Fermi problems in comparative tests allows us to characterize Fermi problems more precisely, as underdetermined modelling tasks from the subject area of arithmetic, that have ambiguous initial states and contain authentic situations.

Fermi problems can be used in comparative tests in Grade 8 and offer a sound method for incorporating authentic situations into test questions, while still satisfying the general criteria of test design. This can allow open contexts and those that are relevant to students, to be incorporated into tests and assessment tasks more frequently. In this way, modelling competencies can be promoted and incorporated into lesson development. Fermi problems can be created with a varying number of magnitudes to be determined by students in several ways. A useful hypothesis on a connection between difficulty levels and number of magnitudes can be proposed and should be evaluated in future projects by undertaking a quantitative study with more Fermi problems designed for assessment. Nevertheless, the formulated hypothesis, in combination with the confirmed definition, could guide task designers through the various issues in creating test items with realistic contexts. Furthermore, task designers should aim at developing Fermi problems that are relevant to test takers in their (later) life.

Acknowledgments

The test items used in this paper were developed and empirically tested under the leadership of the German Institute for Educational Quality Improvement.

Notes

[1] The difficulty level was determined by the German Institute for Quality Development in Education, while piloting the tasks (Greefrath, 2019a). The test items are piloted with an approximately representative sample of students, with each item being attempted by approximately 300 students. Solution frequencies are determined by calculating the percentage of correct solutions among the responses. The difficulty level is then calculated by subtracting the solution frequency from 1, thus a high difficulty level corresponds to a low solution frequency in the pilot study.

References

Albarracín, L., & Gorgorió, N. (2014). Devising a plan to solve Fermi problems involving large numbers. *Educational Studies in Mathematics*, *86*(1), 79–96. https://doi.org/10.1007/s10649-013-9528-9

Bergman Ärlebäck, J. (2009). On the use of realistic Fermi problems for introducing mathematical modelling in school. *The Montana Mathematics Enthusiast*, *6*(3), 331–364.

Bergman Ärlebäck, J., & Albarracín, L. (2019). The use and potential of Fermi problems in the STEM disciplines to support the development of twenty-first century competencies. *ZDM Mathematics Education*, *51*(6), 979–990. https://doi.org/10.1007/s11858-019-01075-3

Bergman Ärlebäck, J., & Bergsten, C. (2010). On the use of realistic Fermi problems in introducing mathematical modelling in upper secondary mathematics. In R. Lesh, P. L. Galbraith, C. R. Haines, & A. Hurford (Eds.), *Modeling students' mathematical modeling mompetencies* (pp. 597–609). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4419-0561-1_52

Blum, W., Drüke-Noe, C., Leiß, D., Wiegand, B., & Jordan, A. (2005). Zur rolle von bildungsstandards für die qualitätsentwicklung im mathematikunterricht. *Zentralblatt für Didaktik der Mathematik*, *37*(4), 267–274. https://doi.org/10.1007/BF02655814

Blum, W., & Leiß, D. (2007). How do students and teachers deal with modelling problems? In C. Haines, P. Galbraith, W. Blum, & S. Khan (Eds.), *Mathematical Modelling (ICTMA 12): Education, Engineering and Economics* (pp. 222–231). Chichester: Horwood. https://doi.org/10.1533/9780857099419.5.221

Blum, W., vom Hofe, R., Jordan, A., & Kleine, M. (2004). Grundvorstellungen als aufgabenanalytisches und diagnostisches Instrument bei PISA. In M. Neubrand (Ed.), *Mathematische kompetenzen von schülerinnen und schülern in Deutschland: Vertiefende analysen im rahmen von PISA 2000* (pp. 145–173). Wiesbaden: VS, Verlag für Sozialwissenschaften.

Büchter, A., & Leuders, T. (2005). *Mathematikaufgaben selbst entwickeln: Lernen fördern - Leistung überprüfen*. Berlin: Cornelsen.

Drüke-Noe, C. (2014). *Aufgabenkultur in klassenarbeiten im fach mathematik*. Wiesbaden: Springer Fachmedien. https://doi.org/10.1007/978-3-658-05351-2

Ehmke, T., Köller, O., & Stanat, P. (2017). Äquivalenz der erfassung mathematischer kompetenzen in PISA 2012 und im IQB-Ländervergleich 2012. *Zeitschrift für Erziehungswissenschaft*, *20*(S2), 37–59. https://doi.org/10.1007/s11618-017-0751-5

Ferrando, I., & Albarracín, L. (2021). Students from grade 2 to grade 10 solving a Fermi problem: Analysis of emerging models. *Mathematics Education Research Journal*, *33*, 61–78 https://doi.org/10.1007/s13394-019-00292-z

Greefrath, G. (2010). Analysis of modeling problem solutions with methods of problem solving. In R. Lesh, P. L. Galbraith, C. R. Haines, & A. Hurford (Eds.), *Modeling students' mathematical modeling competencies* (pp. 265–271). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4419-0561-1_23

Greefrath, G. (2019a). Fermi-aufgaben in vergleichsarbeiten in klasse 8 – Kriterien und ergebnisse. In A. Büchter, M. Glade, R. Herold-Blasius, M. Klinger, F. Schacht, & P. Scherer (Eds.), *Vielfältige zugänge zum nathematikunterricht* (pp. 19–32). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-24292-3_2

Greefrath, G. (2019b). Mathematical Modelling—Background and current projects in Germany. In J. M. Marbán, M. Arce, A. Maroto, J. M. Muñoz-Escolano, & Á. Alsina (Eds.), *Investigación en Educación Matemática XXIII* (pp. 23-41). Valladolid: SEIEM. Retrieved from http://www.seiem.es/docs/actas/23/ActasXXIIISEIEM.pdf

Greefrath, G. (2020). Mathematical modelling competence. Selected current research developments. *Avances de Investigación En Educación Matemática*, *17*, 38–51. https://doi.org/10.35763/aiem.v0i17.303

Greefrath, G., Siller, H.-S., & Ludwig, M. (2017). Modelling problems in German grammar school leaving examinations (Abitur) – Theory and practice. In T. Dooley & G. Gueudet (Eds.), *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education* (pp. 932–939). Dublin: DCU Institute of Education and ERME. Retrieved from https://hal.archives-ouvertes.fr/hal-01933483

Hernandez-Martinez, P., & Vos, P. (2018). "Why do I have to learn this?" A case study on students' experiences of the relevance of mathematical modelling activities. *ZDM Mathematics Education*, *50*(1–2), 245–257. https://doi.org/10.1007/s11858-017-0904-2

Joint Committee on Standards for Educational Evaluation. (2018). *Checklist of the program evaluation standards statements*. Retrieved from https://wmich.edu/evaluation/checklists

Kaiser, G. (2007). Modelling and modelling competencies in school. In C. Haines, P. L. Galbraith, W. Blum, & S. Khan (Eds.), *Mathematical Modelling (ICTMA 12): Education, Engineering and Economics* (pp. 110–119). Chichester: Horwood. https://doi.org/10.1533/9780857099419.3.110

KMK (Ed.). (2012). *Bildungsstandards im fach mathematik für die allgemeine hochschulreife (Beschluss der Kultusministerkonferenz vom 18.10.2012).* Köln: Wolters Kluwer.

Maaß, K. (2006). What are modelling competencies? *ZDM Mathematics Education*, *38*(2), 113–142. https://doi.org/10.1007/BF02655885

Maaß, K. (2010). Classification scheme for modelling tasks. *Journal Für Mathematik-Didaktik*, *31*(2), 285–311. https://doi.org/10.1007/s13138-010-0010-2

Maier, U., Kleinknecht, M., Metz, K., & Bohl, T. (2010). Ein allgemeindidaktisches kategoriensystem zur analyse des kognitiven potenzials von aufgaben. (Fachportal Pädagogik). Retrieved from https://nbn-resolving.org/urn:nbn:de:0111-pedocs-137347

Maier, U., Metz, K., Bohl, T., Kleinknecht, M., & Schymala, M. (2012). Vergleichsarbeiten als instrument der datenbasierten schul- und unterrichtsentwicklung in Gymnasien. In A. Wacker, U. Maier, & J. Wissinger (Eds.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung* (pp. 197–224). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-94183-7_9

Mayring, P. (2014). *Qualitative content analysis: Theoretical foundation, basic procedures and software solution*. Klagenfurt. Retrieved from https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173

Neubrand, M., Klieme, E., Lüdtke, O., & Neubrand, J. (2002). Kompetenzstufen und schwierigkeitsmodelle für den PISA-Test zur mathematischen grundbildung. *Unterrichtswissenschaft*, *30*(2), 100–119.

Niss, M. (1992). Applications and modelling in school mathematics—Directions for future development. In I. Wirszup & R. Streit (Eds.), *Development in school mathematics education around the world* (Vol. 3, pp. 346–361). Reston: NCTM.

Niss, M., & Blum, W. (2020). *The learning and teaching of mathematical modelling*. Abingdon, Oxon ; New York, NY: Routledge.

Niss, M., & Højgaard, T. (2019). Mathematical competencies revisited. *Educational Studies in Mathematics*, *102*(1), 9–28. https://doi.org/10.1007/s10649-019-09903-9

OECD. (2013). *PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD. https://doi.org/10.1787/9789264190511-en

Palm, T. (2007). Features and impact of the authenticity of applied mathematical school tasks. In W. Blum, P. L. Galbraith, H.-W. Henn, & M. Niss (Eds.), *Modelling and applications in mathematics education. The 14th ICMI Study* (pp. 201–208). Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-29822-1_20

Peter-Koop, A. (2004). Fermi problems in primary mathematics classrooms: Pupils' interactive modelling processes. *Mathematics Education for the Third Millennium: Towards 2010. Proceedings of the 27th Annual Conference of the Mathematics Education Research Group of Australasia, Townsville*, *27*, 454–461. Sydney: MERGA.

Peter-Koop, A. (2009). Teaching and understanding mathematical modelling through Fermi-problems. In B. Clarke, B. Grevholm, & R. Millman (Eds.), *Tasks in primary mathematics teacher education* (Vol. 4, pp. 131–146). Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-09669-8_10

Pólya, G. (1981). *Mathematical discovery. On understanding, learning, and teaching problem solving*. New York: John Wiley & Sons Inc.

Prodromou, L. (1995). The backwash effect: From testing to teaching. *ELT Journal*, *49*(1), 13–25. https://doi.org/10.1093/elt/49.1.13

Reit, X.-R., & Ludwig, M. (2015a). An approach to theory based modelling tasks. In G. A. Stillman, W. Blum, & M. Salett Biembengut (Eds.), *Mathematical modelling in education research and practice* (pp. 81–91). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-18272-8_6

Reit, X.-R., & Ludwig, M. (2015b). Thought structures as an instrument to determine the degree of difficulty of modelling tasks. In K. Krainer & N. Vondrová (Eds.), *Proceedings of the Ninth Congress of the European Society for Research in Mathematics Education (CERME9, 4-8 February 2015)* (pp. 917–922). Prague: Charles University in Prague, Faculty of Education and ERME.

Ross, J., & Ross, M. (1986). Fermi problems or how to make the most of what you already know. In H. L. Schoen & M. J. Zweng (Eds.), *Estimation and mental computation* (pp. 175–181). Reston, VA: National Council of Teachers of Mathematics.

Schecker, H., & Parchmann, I. (2007). Standards and competence models: The German situation. In D. Waddington, P. Nentwig, & S. Schanze (Eds.), *Making it comparable. Standards in science education* (pp. 147–164). Münster: Waxmann.

Sriraman, B., & Lesh, R. A. (2006). Modeling conceptions revisited. *ZDM Mathematics Education*, *38*(3), 247–254. https://doi.org/10.1007/BF02652808

van den Heuvel-Panhuizen, M., & Becker, J. (2003). Towards a didactic model for assessment design in mathematics education. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Second international handbook of mathematics education* (pp. 689–716). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-010-0273-8_23

Vos, P. (2018). "How real people really need mathematics in the real world" – Authenticity in mathematics education. *Education Sciences*, *8*(4), 195. https://doi.org/10.3390/educsci8040195