MODESTUM

EURASIA Journal of Mathematics, Science and Technology Education, 2026, 22(8), em2891
ISSN:1305-8223 (online)

OPEN ACCESS

Research Paper

https://doi.org/10.29333/ejmste/19055

Perceived authenticity in solving mathematical reality-based test items: Influencing factors and emotional-motivational effects

Dominik Schlüter [1*], Michael Besser [1], Larissa Altenburger [1], Katrin Rakoczy [2]

[1] Leuphana University of Lüneburg, Lüneburg, GERMANY
[2] Justus Liebig University Giessen, Giessen, GERMANY

Received 09 March 2026 ▪ Accepted 17 July 2026

Abstract

Authenticity is considered a key criterion in mathematical reality-based test items, both to ensure content validity and to promote positive emotions and motivation during processing. However, students' perception of authenticity and its empirical effects remain largely unclear. The study investigates (1) to what extent students perceive aspects of reality-based test items as authentic, (2) to what extent personal characteristics and the occurrence of factual authenticity in reality-based test items predict students' perception of authenticity, and (3) to what extent the perception of authenticity predicts students' emotions and motivations. Data were collected from 258 secondary school students using a paper-and-pencil survey. Results indicate that students perceive the authenticity of aspects in reality-based test items as moderate on average, with substantial variation. Personal characteristics and the occurrence of factual authenticity explain little variance in this perception, whereas perceived authenticity is a highly significant predictor of test-item-specific situational interest, enjoyment, and lower boredom.

Keywords: perceived authenticity, reality-based test items, motivation, emotions, secondary school students, mathematics education

INTRODUCTION

One of the primary goals of mathematics education is to cultivate informed and responsible citizens who can use mathematics to solve real-world problems (Geiger et al., 2015; Niss & Blum, 2020; Siller et al., 2024). In the assessment of this ability using reality-based test items in standardized large-scale tests, authenticity is often considered a key criterion (OECD, 2021, 2023; Tout et al., 2017). In this context, authenticity basically means that the references to reality should be genuine rather than artificially created (Schlüter & Besser, 2024; Vos, 2018). Beyond supporting content validity, authenticity is tentatively assumed to positively affect students' emotions and motivation during assessment (OECD, 2021; National Assessment Governing Board, 2021; Weiss & Müller, 2015). Research into out-of-school learning settings has produced the model of authenticity in learning contexts (Betz et al., 2016; Nachtigall & Rummel, 2021), which offers an explanatory conceptual model for this hypothesis. Here, personal characteristics and varying occurrence of authenticity in learning settings influence the extent to which students perceive the learning setting as authentic. In the model, this perception then affects various variables, such as the students' situational interest. Bearing in mind the potential emotional-motivational effects of authenticity in reality-based test items and this explanatory model, the following research questions (RQ) arise:

(1) To what extent do students perceive aspects of reality-based test items as authentic?

(2) To what extent do personal characteristics and the occurrence of factual authenticity in reality-based test items predict students' perception of authenticity?

(3) To what extent does the perception of authenticity predict students' emotional and motivational effects?

To address these questions, this article first outlines existing considerations and research on authenticity, perceived authenticity and students' perception of test

✉ dominik.schlueter@leuphana.de **(*Correspondence)** ✉ michael.besser@leuphana.de ✉ larissa.altenburger@leuphana.de ✉ katrin.rakoczy@erziehung.uni-giessen.de

Contribution to the literature

- This study contributes to research on real-world connections in mathematics education by examining perceived authenticity as a largely underexplored conceptual perspective on students' engagement with mathematical reality-based test items.
- It extends existing work on authenticity as a design criterion by showing that students' perceived authenticity varies substantially and is only weakly explained by personal characteristics and factual authenticity.
- At the same time, the study provides empirical evidence that perceived authenticity is emotionally and motivationally relevant, as it significantly predicts situational interest, enjoyment, and lower boredom. These results advance current knowledge by highlighting perceived authenticity as an important but insufficiently understood concept in mathematics assessment contexts.

items and their emotional-motivational consequences, before deriving the conceptual model and RQs of the study. The following methods section describes how the RQs were addressed with a quantitative study of students. Next, the results are described and discussed in the context of their interpretations, limitations and implications.

THEORY

Authenticity in Reality-Based Test Items

A key concern of mathematics education is to equip students with the ability to use mathematics in a variety of real-world contexts (Geiger et al., 2015; Niss & Blum, 2020; Siller et al., 2024). This ability is primarily fostered and assessed through dealing with mathematical tasks that address a real-world context, so-called "reality-based tasks" (Knabbe et al., 2025; Leiss et al., 2019). In the context of these tasks, a major concern is that references to reality should correspond to the real world, rather than being artificially invented (Krawitz et al., 2025; Stacey, 2015; Vos, 2018). This idea is reflected in the construct of authenticity. Recognizing that there is always a discrepancy between real-world situations and school mathematical tasks as a whole (Palm, 2008; Vos, 2018), authenticity means basically that an individual task aspect truthfully corresponds with the real world (despite different perspectives existing, see Vos, 2011, and Schlüter & Besser, 2024, for an overview). The events, the question and the use of mathematics are the central aspects of a reality-based task regarding authenticity (Palm, 2009; Schlüter et al., 2026; Siller & Greefrath, 2020). In concrete terms, this means that the described events, the question posed, and the use of mathematics for processing the task should each correspond to real-world events, questions and uses of mathematics. Whether such aspects of a task are 'factually' authentic (a terminology which is clarified further below) is not an inherent property, but rather an attribution in the sense of a social construct (Vos, 2011). In other words, an aspect is considered authentic if there is intersubjective agreement that it corresponds to the real world, bearing in mind that this agreement is subject to cultural and temporal factors.

Authenticity is an important criterion not only for reality-based tasks in the sense of learning tasks, but for reality-based test items in large-scale tests in particular. In both international and national large-scale tests, the importance of authentic references to reality in test items used is explicitly highlighted (Mahler et al., 2020; National Assessment Governing Board, 2021; OECD, 2021, 2023). For instance, in the OECD´s "Programme for International Student Assessment" (PISA) authenticity forms a key criterion when assessing the quality of test items (Stacey, 2015) and the goal of the OECD´s "Programme for the International Assessment of Adult Competencies" (PIAAC) is to implement authentic stimuli and questions in the test items (OECD, 2021). However, the implementation of authentic real-world references in test items is also debated in the literature. For instance, Tout and Spithill (2015) point out that real-world contexts can become too complicated and complex for students, particularly when they require extensive reading. Thus, item developers face the challenge of maintaining authentic aspects while ensuring that the items remain accessible to students in a test situation. There are two main levels of reasons for demanding authenticity in reality-based test items.

1. *Normative claim:* Large-scale tests such as PISA and PIAAC aim to assess the ability to understand and apply mathematics in the real world, in the sense of the concepts of mathematical literacy (OECD, 2023) and numerical literacy (OECD, 2021). In order to ensure content validity, the references to reality in the test items must correspond to the real world, in other words, must be authentic (OECD, 2021, 2023).
2. *Hope for empirical effects:* Secondly, the expectation of emotional-motivational effects is often cited as a reason (Weiss & Müller, 2015). For instance, eTIMSS implements test items "designed to stimulate student motivation through items that simulate real world contexts" (International Association for the Evaluation of Educational Achievement, 2020), PIAAC utilizes authentic

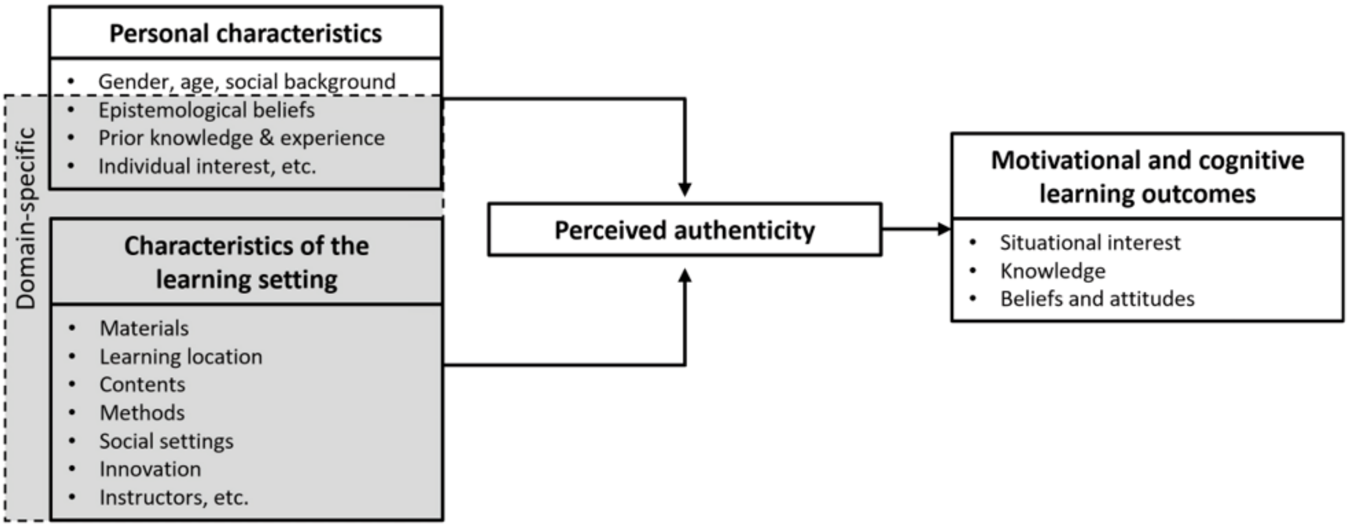

Figure 1. Model of authenticity in learning contexts based on Nachtigall and Rummel (2021), combined with an adaptation (domain-specificity) from Nachtigall et al. (2018a)

situations in its items to "encourage a more positive disposition towards solving relevant and engaging mathematics problems" (OECD, 2021, p. 89) and in PISA the use of real-world references is considered essential in assisting in motivation of students (Stacey, 2015). Similar positions can also be found in national large-scale tests (Mahler et al., 2020; National Assessment Governing Board, 2021).

Such assumptions regarding empirical effects on emotional-motivational experiences are supported by initial findings of studies investigating the impact of authenticity in reality-based learning tasks. For instance, Vos et al. (2007) and Lowrie (2011) report in qualitative studies that the use of authentic objects can have a positive impact on motivation. In another qualitative study, Tran et al. (2019) found that students showed a more positive disposition towards a more-authentic task in contrast to a less-authentic one. Hartmann and Schukajlow (2021) hypothesize that the authenticity of problems may be important for the development of emotions and interest, but this remains to be tested in future studies. However, the role authenticity actually plays in processing reality-based test items is largely unclear, even though it is a key feature of many prominent large-scale tests.

Perceived Authenticity as a Link Between Intended Authenticity and Emotional-Motivational Experiences

To explain assumed effects of intended authenticity on emotional-motivational variables, a conceptual distinction is necessary between *factual authenticity* and *perceived authenticity* (based on Betz et al., 2016; Hohrath et al., 2024; Nachtigall & Rummel, 2021; Weiss & Müller, 2015). *Factual authenticity* is an item-level characteristic that refers to whether aspects of a test item can be intersubjectively justified as corresponding to the real world. The term 'factual' does not imply that authenticity is an objective property of test item aspects in an absolute sense. Rather, it denotes the item-level attribution described before. By contrast, *perceived authenticity* is conceptualized as a student-by-item appraisal. It refers to the extent to which students perceive aspects of a test item as corresponding to the real world. From a theoretical perspective, factual and perceived authenticity should not be assumed to correspond automatically, as students may recognize, interpret, or evaluate the relevant item characteristics differently. Although the concept of perceived authenticity has received less focus in mathematics education research yet, particularly with regard to reality-based test items, there are existing concepts that are primarily used in the context of out-of-school learning settings in various disciplines (e.g., Betz et al., 2016; Nachtigall et al., 2022; Schüttler et al., 2021; Stamer et al., 2021).

In this area, Betz et al. (2016) developed a central impact model, which has since been further developed and adapted in various ways (see Figure 1 for the well-established extension by Nachtigall & Rummel, 2021). This model has already been used in learning settings in various subject areas (history: Zachrich et al., 2024; linguistics: Betz, 2018; social sciences: Nachtigall et al., 2018b). In the model, there are two input boxes: one for personal characteristics (such as demographic variables, individual interests or prior knowledge, see Betz et al., 2016; Nachtigall & Rummel, 2021) and one for characteristics of the learning setting (in terms of *factual authenticity* of central aspects, e.g., the location if it corresponds to a real research location, the data if it corresponds to real data or the method if it corresponds to a real scientific method, see Finger et al., 2022; Nachtigall & Rummel, 2021). Nachtigall et al. (2018a) point out that these two input components should be considered within the context of their specific domain (see grey box in **Figure 1**). These two input components then affect the perception of authenticity (Betz et al., 2016). Unlike the input component at the bottom left, this does not refer to the factual authenticity of certain aspects of the learning setting, but to the extent to which learners themselves perceive the authenticity of these aspects (Finger et al., 2022). This perception then has a subsequent effect on learners. Besides cognitive effects, motivational and emotional effects on learners during processing the learning setting are often cited (Nachtigall et al., 2022). The main focus here is on situational interest (Nachtigall & Rummel, 2021).

Empirical findings on this model in the field of out-of-school learning settings are mixed. Some studies suggest that factual authenticity can increase perceived authenticity, whereas others do not find such a direct relationship (Gulikers et al., 2005; Nachtigall et al., 2018b; Nachtigall & Rummel, 2021; Schüttler et al., 2021; Stamer et al., 2021). The connection between perceived authenticity and emotional-motivational variables, particularly situational interest, has been demonstrated (Betz, 2018; Nachtigall & Rummel, 2021; Nachtigall et al., 2018b; Schüttler et al., 2021).

Emotional-Motivational Consequences of Students' Perceptions of Tasks and Test Items

Research highlights that students' perceptions of tasks and test items have emotional and motivational consequences (Dowley & Rice, 2022; Tanaka & Murayama, 2014). In the area of research on reality-based tasks and test items, the following three emotional-motivational constructs are considered particularly important (Hartmann & Schukajlow, 2021; Kanefke & Schukajlow, 2024). On a motivational level, the focus is on the construct of situational interest, which refers to an unstable motivational state of focused attention and affective reaction (Hidi & Renninger, 2006). On an emotional level, the main focus is on the constructs of enjoyment and boredom. Both emotions can be characterized in the context of the control-value theory of achievement emotions (Pekrun, 2006). According to this theory, two appraisals are central to the arousal of emotions: control (e.g., do I think I am capable of solving the test item?) and value (how valuable is the test item to me?). Enjoyment is a positive and activating emotion that occurs when the appraisal of control and value is high. Boredom is a negative and deactivating emotion that occurs when the appraisal of control is low or very high and the appraisal of value is low (Pekrun, 2006).

Although mathematical large-scale tests hark back to authenticity, assuming it has positive emotional-motivational effects, the extent to which students actually perceive authenticity, the factors influencing this perception and the effects of these perceptions remain an area where empirical research is still lacking but needed in mathematics education (Schlüter & Besser, 2024; Vos, 2018). One study focusing on the perception of authenticity in reality-based test items originating from a discipline closely related to mathematics can be found in Weiss and Müller (2015). Among other things, the study used a set of PISA physical science test items to analyze the extent to which students perceive the test items as authentic and interesting. Contrary to the basic assumption of PISA, it was shown that both perceived authenticity and interest are at best medium (Weiss & Müller, 2015). However, neither influencing factors on the perception of authenticity nor the impact of perceived authenticity on emotional-motivational experiences were assessed. Sundtjønn (2025) further demonstrated that students can form judgements about the authenticity of aspects of a vocational learning task, and that aspects that are not perceived as authentic can be demotivating.

CONCEPTUAL MODEL AND RESEARCH QUESTIONS

To address the research gap outlined previously, the impact model from research on out-of-school learning settings (Betz et al., 2016; Nachtigall & Rummel, 2021) has been adapted to the context of students working on mathematical reality-based test items (see **Figure 2**).

The input component on the top left contains socio-demographic and domain-specific variables. The domain-specific variables are those that are considered central to the processing of mathematical tasks and test items. These include general emotional-motivational variables related to mathematics (Schukajlow et al., 2022; Živković et al., 2023), general language abilities (Leiss et al., 2019; Prediger et al., 2018) and non-verbal cognitive abilities (Amland et al., 2025; Bahar & Maker, 2015) in particular. The input component in the bottom left has also been adapted so that it now includes the factual authenticity of reality-based test items. As in Betz et al. (2016), these two components affect students' perceptions of authenticity. In line with the outlined research interest, the box on the right contains the emotional and motivational effects on the students, including situational interest, enjoyment and boredom.

The first RQ focuses on perceived authenticity, which is the central construct of the present study and is located at the center of the conceptual model. To examine this construct more closely, the first RQ is:

RQ1: To what extent do students perceive aspects of reality-based test items as authentic?

The two input components on the left side of the model, personal characteristics and factual authenticity, then lead to the question of whether these variables predict students' perceived authenticity:

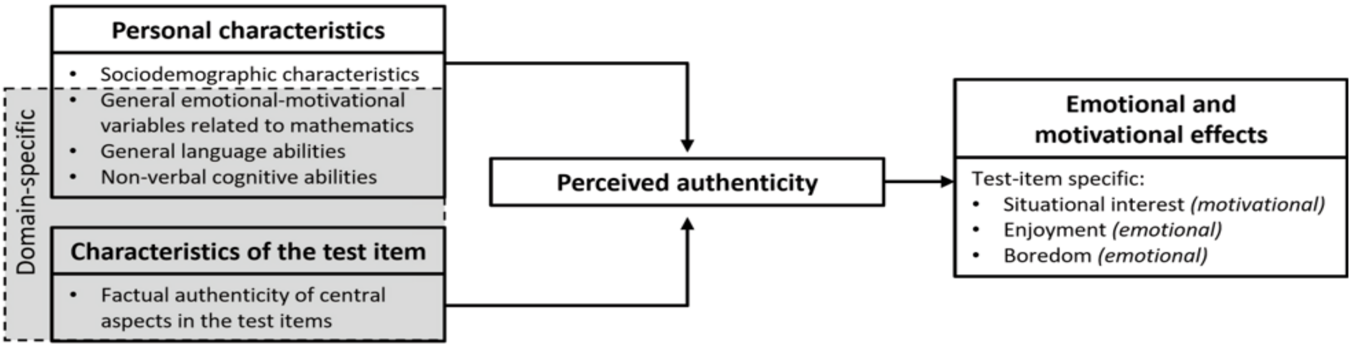

Figure 2. Conceptual model of the study (the authors' own elaboration)

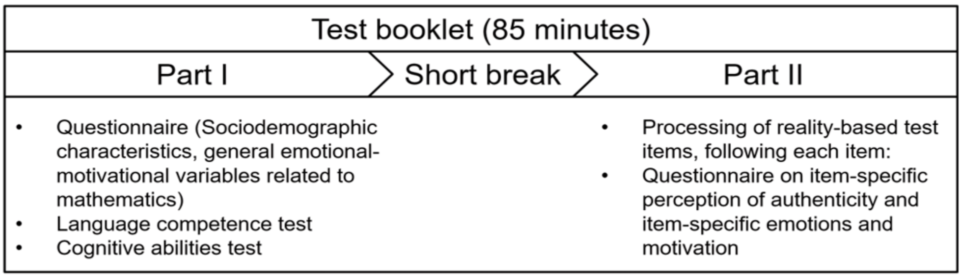

Figure 3. Study procedure (the authors' own elaboration)

RQ2: To what extent do personal characteristics and the occurrence of factual authenticity in reality-based test items predict students' perception of authenticity?

Finally, the outcome component of the model concerns the assumed emotional-motivational effects of perceived authenticity during test item processing. This leads to the third RQ:

RQ3: To what extent does the perception of authenticity predict students' situational interest, enjoyment, and boredom during test item processing?

METHODS

Sample

In the first quarter of 2025, a convenience sample of 258 eighth graders (female: 49.2%, male: 50.8%, diverse: 0%) from 14 classes at five German secondary schools participated in this study. The participating schools included Realschulen, a type of German secondary school typically leading to an intermediate school-leaving qualification, as well as various forms of comprehensive secondary schools offering pathways to different school-leaving qualifications. The schools were located in the German federal states of Schleswig-Holstein and Lower Saxony. The assessment was conducted in regular school contexts. The mean (M) age of the students in the sample was 13.7 years (standard deviation [SD] = .69) and 29.8% reported a migration background. In terms of socio-economic status, the home educational resources can be classified as average (HEDRES index of M = 1.21, SD = 0.20; for comparison the German sample in PISA 2018 had M = 1.18, SD = 0.18, see Mang et al., 2021).

Study Design

The present study is a cross-sectional study in which students completed a test booklet once under supervision. The whole testing lasted 85 minutes and was conducted by trained test administrators. The testing consisted of two parts (see Figure 3) and the order in which the constructs were assessed corresponded to the order in the conceptual model. The first part consisted of three components: first, a questionnaire on socio-demographic variables as well as emotions and motivations related to general mathematics education; second, a test to assess language skills, and third, a test of general cognitive abilities. This was followed by a 5-minute break to allow the students to recover. In the second part, the students processed reality-based test items. After each test item, they completed a survey with questions specific to the test item. The aim was to assess their perception of authenticity as well as emotions and motivation.

Description of Reality-Based Test Items Used

For part II of the assessment, students worked on twelve reality-based test items. These were taken from the German national standardized test VERA 8 ("Vergleichsarbeiten der Jahrgangsstufe 8"). The test is designed to assess mathematical competencies of students in the eighth grade (Eckhardt, 2021). The item development process includes expert review and piloting, among other steps, to ensure that items are of sufficient empirical quality and are appropriate from a didactic perspective. To ensure that the selected items are reality-based test items, only those test items that explicitly address a real-world context were chosen. In

Finder's fee

In Germany, the finder's fee is 5% of the value of lost property worth up to €500. If the value of the lost property is over 500€, it is calculated as follows: the finder's fee for €500 is increased by 3% of the value exceeding €500.

Source: § 971 German Civil Code

Janine found an envelope containing €2,000 in cash and handed it in to the police.

How much is her finder's fee?

Mark with a cross:

☐ 60€ ☐ 70€ ☐ 100€ ☐ 160€

Figure 4. Test item Finder´s Fee (2019, M16524), translated by the authors, copyright to the original German texts/images/audio/graphics: IQB e. V., Licence: Creative Commons (CC BY). Full text available at: https://creativecommons.org/licenses/by/3.0/de/legalcode

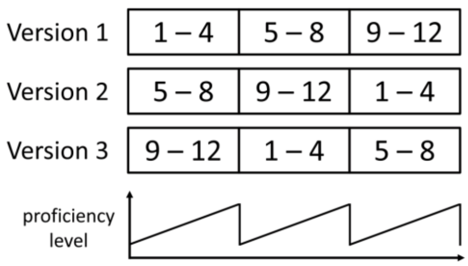

Figure 5. Arrangement of test items (the authors' own elaboration)

these test items, each item consists of a stimulus and a corresponding question that must be solved. The stimulus usually consists of an introductory text referring to a real-world context and the question requires a solution process in which not only intra-mathematical content is relevant, but also information from the real-world context. An example test item is provided in **Figure 4**. In this item, the stimulus introduces the legally regulated concept of a finder's fee in Germany and presents a concrete situation together with a corresponding question. To answer the question, students must translate between the real-world situation and mathematics, even though the required modelling activity is relatively limited. They must identify the relevant information from the reality-based situation, relate it to the rule described in the stimulus, and apply the corresponding mathematical procedure to determine the finder's fee. When selecting the test items, care was taken to ensure variance in terms of factual authenticity (see instruments).

To reduce fatigue and increasing frustration leading to improper test item processing, the test items were arranged in an increasing difficulty block structure (Drüke-Noe, 2012). There are three recurring cycles of difficulty, with each cycle beginning with an easy item and ending with a difficult one (see **Figure 5**). The ranking was based on the proficiency level categorization issued by the test developer, which was formed based on empirical difficulty. To eliminate possible order effects, the three task blocks were counterbalanced in the test booklets.

Instruments

Appendix A lists all instruments used in the study, along with the respective items, in their original German form and in English translation. The following section provides a more in-depth description of these instruments.

Personal characteristics

Socio-demographic characteristics: The first part of the test booklet contained questions on socio-demographic variables including age, gender, migration background and educational resources at home. These variables were assessed using established items from large-scale assessments. Gender was assessed using a three-option item: male, female, or diverse. Since diverse was not selected, the variable is treated dichotomously (female = 1, male = 2). The student's migration background was assessed using three items relating to their own origin, and those of their mother and father taken from PISA 2018 (Mang et al., 2021). These items were used to form the following two dichotomous categories: students without a migration background = 1, students with a migration background = 2. In terms of the socio-economic status, the students completed items from the PISA 2018 background questionnaire on the availability of household resources (Mang et al., 2021). The HEDRES (Home Educational Resources) index was subsequently calculated using seven items related to possessions relevant to learning (internal consistency $\alpha = .50$, $\omega = .51$, comparable to PISA with $\alpha = .56$, see Mang et al., 2021). The index ranges from 1 to 2 and is calculated from the average rating value, with resource available = 1 and resource not available = 2. In this exceptional case, this means that a higher value is associated with a lower level of home educational resources.

General emotional-motivational variables related to mathematics: Mathematics-related interest, mathematics-related self-efficacy, mathematics-related self-concept and mathematics anxiety are considered to be key general emotional and motivational constructs in mathematics education (Schukajlow et al., 2022, 2023; Živković et al., 2023). These constructs were assessed using established questionnaire scales from previous research in mathematics education. To assess mathematics-related interest, the PALMA mathematics interest scale was used (taken from Frenzel et al., 2012). The scale consists of six items (α = .80, ω= .81, example item: "After a math class, I am often curious about what we are going to do in the next lesson."). The other three constructs were assessed using established scales provided by Ramm et al. (2012). The mathematics-related self-efficacy scale consists of four items (α = .82, ω= .82, example item: "I am confident that I can achieve good results in mathematics homework and exams."), the mathematics-related self-concept scale consists of five items (α = .88, ω= .88, example item: "I learn maths quickly.") and the mathematics anxiety scale consists of five items (α = .80, ω= .81, example item: "I feel very tense when I have to do my mathematics homework."). All items relating to these four constructs were assessed in a mixed order using a four-point rating scale (1 = does not apply at all to 4 = fully applies).

Language competence: To assess students' language competence, a brief cloze test was implemented. In this test format, students are presented with short texts from which the second part of certain words has been deleted and students must complete these. This test format is well established in educational research, where it is considered to be a reliable and valid instrument for measuring general language competence (Grotjahn, 2010). This study used the standardized DCLL+3 test[1], which is designed for grades 7 and 8. Two short texts were implemented, each with 25 gaps to fill in. A student's total score for language competence was derived based on correctness of their gap completions, with a maximum possible score of 100. Two trained coders rated the correctness of the gap fillings, achieving high agreement in a consistency test based on 20 samples (Cohen's $\varkappa$ = 0.94, $p<.001$).

Non-verbal cognitive abilities: Language-independent cognitive abilities were assessed using the standardized German cognitive ability test KFT 4-12+R (Heller & Perleth, 2000). Here, form A of the non-verbal N-test 2, designed for grade 8, was used. The test comprises 25 items, in each of which a second pair of figures with the same relationship characteristics must be formed from a given pair of figures. One of the figures in this second pair has already been given to the students. The matching second figure must be selected from five possible answers in a single-choice format. The final score was derived as the sum of correctly answered items, with a maximum possible score of 25.

Factual authenticity in reality-based test items

The factual authenticity of each test item was assessed in accordance with the model for authenticity in modelling test items (MAMTI) (Schlüter & Besser, 2024). Accordingly, based on Vos' (2018) concept of authenticity, individual aspects of the test items are assessed separately and can be either authentic or non-authentic. These aspects include the three central aspects of mathematical reality-based test items: the events described in the item, the question posed, and the use of mathematics required to solve the item. An aspect is considered authentic if it can be shown to be genuinely derived from reality. The raters had access to item-related background materials and used standardized online searches to support this evaluation. The test item analysis by Schlüter et al. (2026) indicates that authenticity occurs in recurring profiles across test item aspects. Test items with similar patterns of authentic and non-authentic aspects could be grouped into different types based on the aspect-wise coding. The vast majority of the test items fell into the following three types:

- Type 1: "Few authentic aspects", which is characterized in particular by the fact that the events, the questions and the use of mathematics are not authentic.
- Type 2: "Authentic events, objects & data", in which the events are authentic, but the question and the use of mathematics are not.
- Type 3: "Plenty authentic aspects", in which the events, the question and the use of mathematics are authentic.

The twelve reality-based test items used in the study can be assigned to these three types: two test items to type 1 ("few authentic aspects"), five test items to type 2 ("authentic events, objects & data") and five to type 3 ("plenty authentic aspects"). The unequal distribution of test items across types resulted from the existing item pool and from the fact that, when selecting the test items, care was also taken to ensure that they covered a broad range of content areas and task requirements. The unequal distribution did not preclude the use of the linear mixed-model approach described later in the article, in which test items were accounted for as random effects. The coding of the authenticity of the individual test item aspects and their subsequent assignment to one of the three types draw on the test item analysis by Schlüter et al. (2026). In this analysis, two independent raters assessed the authenticity according to the MAMTI (Schlüter & Besser, 2024), with a significant agreement at

[1] The DCLL+3 test is available from the Test Development and Diagnostics Department of the Hamburg Institute for Educational Monitoring and Quality Assurance.

Table 1. Correlation of study variables (n = 258)

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Age	13.72	0.69	-												
2. Gender	1.51	0.50	.01	-											
3. Migration background	1.30	0.46	.13*	-.00	-										
4. Educational resources at home	1.21	0.20	.07	.10	.18**										
5. Math.-related interest	2.04	0.61	.07	.03	.07	.05									
6. Math.-related self-efficacy	2.45	0.64	-.01	.14*	-.07	.03	.61**								
7. Math.-related self-concept	2.38	0.74	.05	.13*	-.04	.01	.60**	.76**							
8. Math. anxiety	2.38	0.67	-.02	-.15*	.09	.15*	-.33**	-.49**	-.60**						
9. Language competence	64.42	18.39	-.12	-.04	-.40**	-.22**	-.15*	-.03	-.07	-.09					
10. Non-verbal cognitive abilities	12.92	6.63	-.19**	-.13*	-.08	-.08	.11	.19**	.28**	-.33**	.21**				
11. Test item-specific perc. auth.	2.63	0.40	-.08	-.04	-.06	-.07	.06	.14*	.06	.01	.11	.13*			
12. Test item-specific sit. interest	2.38	0.57	-.03	.10	.05	-.01	.37**	.17**	.11	-.08	-.01	.08	.42**		
13. Test item-specific boredom	2.37	0.60	-.06	-.08	.05	.08	-.27**	-.16*	-.12	.16*	.02	-.04	-.32**	-.69**	
14. Test item-specific enjoyment	2.28	0.59	-.03	.14*	.04	-.06	.37**	.13*	.08	-.06	-.05	.08	.38**	.87**	-.67**

Note. Gender: 1: Female & 2: Male; Migration background: 1: Without migration background & 2: With migration background; the correlations between dichotomous and metric variables correspond to point biserial coefficients and correlations between dichotomous variables correspond to phi coefficients; * indicates p < .05 and ** indicates p < .01

a substantial level (Cohen's $\varkappa = 0.69$, $p < .001$; percentage agreement 0.86). In summary, the variable factual authenticity thus has three categorical values, corresponding to the three item types.

Test item-specific perceived authenticity

The students were asked to indicate their perception of authenticity in the test item by rating statements that were presented immediately after each task. As there is no established instrument for assessing students' perceived authenticity of test items in mathematical reality-based tests, a new scale was developed for this study. The item development was theory-driven and based on the three central aspects of authenticity in mathematical reality-based test items: the events, the questions, and the use of mathematics. Accordingly, a four-point rating scale (1 = does not apply at all to 4 = fully applies) was used to assess agreement with the following three items:

(1) "The introduction states that [description of the event(s)]. This event/these events is/are based on real life."

(2) "The question posed in the task is one that arises in real life as well."

(3) "The mathematical approach to solving this problem is the same as in real life."

Prior to the main study, the items were piloted to examine whether students understood the wording, the intended interpretation of the items, and the response format. A scale was created for the test item-specific perceived authenticity from these items in a long data format. The scale showed moderate internal consistency, with $\alpha = .66$ and $\omega = .67$. Because the scale comprised only three items, Cronbach's alpha should be interpreted with caution (Clark & Watson, 2019). The average inter-item correlation was .39, which falls within the range recommended by Clark and Watson (2019), indicating an adequate degree of interrelatedness between the items.

Test item-specific emotions and motivation

The assessment of situational interest, enjoyment and boredom for each test item was carried out immediately after the previous items concerning perceived authenticity were completed. Single-item measures were used for each construct in order to reduce response and reflection time (Kanefke & Schukajlow, 2024). The item for situational interest was "I found working on the task interesting." The item for enjoyment was "I enjoyed working on the task." For boredom, it was "I was bored working on the task." These items were minimal adaptations of Kanefke and Schukajlow (2024), who used comparable task-specific single-item measures based on well-evaluated scales from previous studies. The items were assessed using a 4-point rating scale (1 = does not apply at all to 4 = fully applies).

Table 1 shows the correlations between the variables used in this study. To do this, the test item-specific variables were converted into a scale across all test items.

Analysis

The proportion of missing data in the dataset is 0.92%, which is relatively low for empirical educational research (Enders, 2003). Missing values tend to be unsystematic, for instance in cases of unanswered items or inadmissible information. For specific variables, the percentage of missing values fluctuates between 0% and 3.9%. Missing values were handled with multiple imputations by chained equations using the R package 'mice'.

To answer RQ1, the mean values for each of the three single items (perceived authenticity of the events, the question and the use of mathematics) are reported, as well as the resulting overall scale for perceived authenticity, for each reality-based test item and across all reality-based test items. RQ2 was addressed by estimating cross-classified linear mixed models (LMMs), with perceived authenticity as the dependent variable. This modelling approach was chosen because of the

Table 2. *M* and *SD* values of perceived authenticity

Test item	Type	Perceived authenticity			
		Events	Question	Use of math.	Overall scale
Radio license fee	1	2.40 (0.89)	2.25 (0.87)	2.48 (0.81)	2.38 (0.70)
Advertising lottery	1	2.52 (0.87)	2.32 (0.87)	2.29 (0.83)	2.38 (0.67)
Fountain	2	2.44 (0.90)	1.94 (0.88)	2.26 (0.89)	2.21 (0.70)
Lake Baikal	2	2.78 (0.82)	2.15 (0.87)	2.43 (0.80)	2.45 (0.62)
Right-hand bend	2	3.22 (0.75)	2.22 (0.87)	2.37 (0.83)	2.60 (0.55)
Keys	2	2.96 (0.85)	2.48 (0.93)	2.34 (0.89)	2.60 (0.64)
Croissant	2	3.31 (0.75)	2.84 (0.85)	2.82 (0.86)	2.99 (0.60)
Board game Ludo	3	3.14 (0.88)	2.38 (0.93)	2.30 (0.90)	2.61 (0.69)
Guessing in a test	3	3.00 (0.83)	2.68 (0.87)	2.52 (0.85)	2.73 (0.67)
Finder’s fee	3	2.93 (0.79)	2.69 (0.84)	2.74 (0.74)	2.79 (0.62)
Snow depth	3	3.01 (0.75)	2.70 (0.81)	2.79 (0.82)	2.83 (0.62)
Mobile phone fee	3	3.01 (0.74)	2.97 (0.72)	2.87 (0.76)	2.95 (0.57)
All items		**2.89 (0.87)**	**2.47 (0.91)**	**2.52 (0.86)**	**2.63 (0.68)**

crossed structure of the data: Each observation represented the responses of one student to one of the twelve reality-based test items (long data structure with 3096 observations). Thus, observations were not independent because responses were cross-classified by students and test items. To control for the individual variability of the students and the test items, both were treated as random effects. In accordance with the conceptual model in **Figure 2**, personal characteristics (age, gender, migration background, home educational resources, mathematics-related interest, mathematics-related self-efficacy, mathematics-related self-concept, mathematics anxiety, language competence, non-verbal cognitive abilities) and test item-related characteristics (factual authenticity) were included in the model as fixed effects. To examine the relative contributions of student- and item-related characteristics, the analysis first specified a model with only personal characteristics, then a model with only test item-related characteristics, and finally a full model including both sets of predictors. The assumptions of LMMs, including linearity, normally distributed residuals and random effects, homoscedasticity, and conditional independence, were considered. Although there are high correlations in some cases, there is no expectation of problematic multicollinearity for the model, as all VIF values of the IVs are below 3.2. All metric predictors were z-standardized prior to analysis. For RQ3, three separate cross-classified LMMs were estimated to investigate the effect of perceived authenticity on students’ test item-specific situational interest, enjoyment and boredom. The statistical modelling approach was similar to that used in the previous analysis. This time, perceived authenticity was included as a standardized fixed effect predictor, with test item-specific situational interest, enjoyment, and boredom serving as the dependent variables in the individual models. Separate models were estimated because situational interest, enjoyment, and boredom represent related but conceptually distinct emotional-motivational outcomes.

RESULTS

Research Question 1

With regard to RQ1, **Table 2** shows that the overall mean value of perceived authenticity across all test items is moderate (2.63 on a scale of 1 to 4). The differences in perception between the test items are already apparent. For instance, the mean score for the ‘Fountain’ test item is 2.21, whereas the mean score for the ‘Croissant’ test item is 2.99. Moreover, looking at the scores for each item on the perceived authenticity scale shows that the *SD* is typically around 0.9. With a rating scale with a span of 3, the maximum possible *SD* is 1.5. This means that the *SD* in this case is 60% of the maximum, which is considered very high. Consequently, it can be concluded that students’ perception of authenticity is marked by great heterogeneity.

Research Question 2

Three LMMs were estimated for RQ2 (see **Table 3**): a model with only personal characteristics as predictors (model 1), a model with only item-related characteristics as predictors (model 2) and a full model including both sets of predictors (model 3). Overall, the regression coefficients of the predictors remain stable across the three models. Despite the large number of predictors included in the model, the fixed effects only explain a small proportion of the variance. The personal characteristics model achieves an R^2 value of 0.022, adding item-related characteristics increases this to 0.067. Of the person-related variables, only general mathematics-related self-efficacy was found to be a significant predictor of perceived authenticity ($\beta = 0.098$, SE = 0.041, $p = .017$). In terms of item-related characteristics, a significant effect is observed when comparing type 1 (“few authentic aspects”) with type 3 (“plenty authentic aspects”). After applying post-hoc pairwise comparisons with a Tukey correction for multiple testing, the effect remained only at the margin of statistical significance (see **Table 4**).

Table 3. Effects on the student´s perceived authenticity

Measure	Model 1		Model 2		Model 3	
	β	SE	β	SE	β	SE
Fixed effects						
Intercept	**2.64******	0.08	**2.38******	0.15	**2.39******	0.15
Age	-0.01	0.03	-	-	-0.01	0.03
Gender	-0.02	0.05	-	-	-0.02	0.05
Migration background	-0.00	0.06	-	-	-0.00	0.06
Home educational resources	-0.03	0.03	-	-	-0.03	0.03
Mathematics-related interest	0.00	0.03	-	-	0.00	0.03
Mathematics-related self-efficacy	**0.10***	0.04	-	-	**0.10***	0.04
Mathematics-related self-concept	-0.03	0.04	-	-	-0.03	0.04
Mathematics anxiety	0.06	0.03	-	-	0.06	0.03
Language competence	0.03	0.03	-	-	0.03	0.03
Cognitive abilities	0.05	0.03	-	-	0.05	0.03
Factual authenticity (T1 vs. T2)	-	-	0.19	0.17	0.19	0.17
Factual authenticity (T1 vs. T3)	-	-	**0.40***	0.17	**0.40***	0.17
Random effects						
Student SD	0.369		0.375		0.369	
Test item SD	0.240		0.204		0.204	
Residual SD	0.518		0.518		0.518	
Model statistics						
Marginal R^2	0.022		0.046		0.067	
Conditional R^2	0.432		0.432		0.439	
Deviance	5,252.48		5,263.56		5,246.36	
AIC	5,280.48		5,275.56		5,278.36	
BIC	5,365.01		5,311.79		5,374.96	

Note. Estimates and R^2 are REML-based; fit indices and likelihood-ratio tests are ML-based; *p < 0.05; **p < 0.01; ***p < 0.001

Table 4. Post hoc analyses on factual authenticity

Comparisons of group means		
Type	Estimated marginal means	SE
Type 1 (few authentic aspects)	2.38	0.15
Type 2 (authentic events, objects & data)	2.57	0.10
Type 3 (plenty authentic aspects)	2.78	0.10
Pairwise comparison (with Tukey correction for multiple testing)		
Type comparison	p	
Type 1 vs. type 2	0.504	
Type 1 vs. type 3	0.051	
Type 2 vs. type 3	0.240	

Model comparisons show that model 3 outperforms model 1 (χ^2 (2) = 6.13, p = .047). In comparison to model 2, model 3 only offers a slight but not statistically significant additional benefit (χ^2 (10) = 17.20, p = .070). However, a key finding emerges clearly in the end: Fixed effects account for only a small proportion despite many constructs included (0.07), whereas the contribution associated with the random effects is substantially larger (0.37), but there is still a great deal of unexplained variation even within students and tasks (see high residual variation).

Research Question 3

Three separate LMMs were estimated to investigate the effect of perceived authenticity on students' test item-specific situational interest, enjoyment and boredom. As shown in **Table 5**, which reports the models relevant to RQ3, perceived authenticity is a highly significant predictor in all three models. The effects remain highly significant even after correction for multiple testing using the Holm method (all p < .001). This demonstrates that a higher perception of authenticity results in increased situational interest, enhanced enjoyment and reduced boredom. The explanation of variance differs: perceived authenticity explains over 10% of the variance in situational interest, around 7% in enjoyment and around 5% in the case of boredom. It should be noted that situational interest, enjoyment, and boredom were highly correlated with each other (see **Table 1**). This pattern suggests that perceived authenticity elicits more than one affective response and consistently shapes a broader emotional-motivational profile during test item processing.

Table 5. Effects on the student´s situational interest, enjoyment and boredom

Measure	Dependent variable: situational interest		Dependent variable: enjoyment		Dependent variable: boredom	
	β	SE	β	SE	β	SE
Fixed effects						
Intercept	**2.38******	0.06	**2.28******	0.07	**2.38******	0.05
Perceived authenticity	**0.29******	0.02	**0.25******	0.02	**−0.21******	0.02
Random effects						
Student SD	0.48		0.51		0.54	
Test item SD	0.19		0.21		0.13	
Residual SD	0.69		0.69		0.72	
Model statistics						
Marginal R^2	.103		.074		.052	
Conditional R^2	.422		.435		.402	

Note. *p < 0.05; **p < 0.01; ***p < 0.001

DISCUSSION

Summary and Discussion of the Results of Research Questions 1 to 3

The results suggest that perceptions of authenticity vary considerably between students as well as between test items and the predictors under investigation did hardly help to explain how these perceptions are formed. Thus, perceived authenticity should not be understood merely as a direct reflection of the item-level characteristic of factual authenticity. Rather, it appears to be a broader appraisal construct. Possible explanations for the observed heterogeneity include students´ prior experiences with the respective contexts, their familiarity with the situations described in the items, and individual meaning-making processes during test item processing (Hohrath et al., 2024). Also, item-specific characteristics not captured in the present study may have contributed to the differences between items. This interpretation is also consistent with the finding that the variance components at the student and item levels were substantially larger than the variance explained by the fixed effects included in the model. Thus, perceived authenticity may be shaped by additional factors that were not included in the present conceptual model. The remaining unexplained variation further underlines the need for future research to investigate more closely how students' perceptions of authenticity are formed. It was subsequently demonstrated that perceived authenticity is a highly significant predictor of situational interest, enjoyment, and the reduction of boredom. A possible explanation is that perceived authenticity makes the mathematics in a test item seem more meaningful and relevant to students. In this way, students may be more likely to take the real-world reference seriously and engage in sense-making rather than treating the context as merely decorative (Palm, 2008; Verschaffel et al., 2000), which may foster situational interest and enjoyment and reduce boredom. For situational interest, perceived authenticity accounted for approximately 10% of the variance. Given that situational interest is shaped by multiple individual and contextual factors (Knogler et al., 2015), the variance explained by a single test item-specific predictor can be regarded as non-trivial. This result highlights the potential educational relevance of perceived authenticity for students' engagement with reality-based test items.

Limitations

When interpreting the results, the limitations of the study must be considered. In the study design, students first reported their perception of authenticity, followed by questions on emotions and motivational states. While this sequence was chosen deliberately to reflect the conceptual model, it remains possible that the order of measurement may have influenced how students responded to the subsequent items. In addition, the study takes an initial look at the effects of factual authenticity with existing test item types. This could be examined in more detail in future studies by manipulating and considering the three aspects of events, questions and use of mathematics individually. Here, another limitation is that test item type 1 was represented by only two reality-based test items. Although this does not preclude estimation within the cross-classified LMM framework, it limits the generalizability of comparisons involving this item type. The limited generalizability must be considered in greater depth overall, as the study involved a specific test setting in which a convenience sample was used to work on twelve large-scale test items. The extent to which the results can be transferred to other cultural contexts, age groups, or forms of reality-based test items is largely unclear and the task of future research.

Implications

(1) As a first contribution, the present results allow initial conclusions to be drawn about the underlying model from out-of-school learning settings (Betz et al., 2016; Nachtigall & Rummel, 2021) and its transferability to the context of mathematical reality-based test items. Nachtigall

et al. (2022) describe authenticity as a 'double-edged sword', a view that is also reflected by studies in which factual authenticity leads to a higher perception of authenticity (Schüttler et al., 2021; Stamer et al., 2021), and others in which it does not (Gulikers et al., 2005; Nachtigall & Rummel, 2021). In line with this, the present study finds an effect at the significance threshold without a major impact on the perception of authenticity. However, there are arguments that support the value of the model for researching reality-based test items in mathematics. Firstly, the regression coefficients for the predictors remain stable across the three models in RQ2. Although this may partly be due to the stable estimation of LMMs with random effects, it nevertheless suggests that inputbox 1 and inputbox 2 may be separately associated with perceived authenticity, as in the model. Secondly, perceived authenticity is clearly confirmed as a highly significant predictor of emotional-motivational variables.

(2) Moreover, the study's results expand the theoretical understanding of emotional and motivational processes when students encounter reality-based situations. Previous research on this topic has focused on various aspects, including the emotional and motivational experiences when processing tasks with or without a reference to reality (Schukajlow et al., 2012), whether reality-based tasks are open or closed (Quarder et al., 2025; Schukajlow et al., 2022), and whether such tasks are completed inside or outside the classroom (Hartmann & Schukajlow, 2021). Contrary to theoretical expectations, research has shown that students do not generally find tasks with real-life relevance more interesting or enjoyable (Krug & Schukajlow, 2013; Schukajlow et al., 2012). Hartmann and Schukajlow (2021) suggest that the authenticity of problems may be important for the development of emotions and interest, but this remained to be tested in future studies. In this regard, the study findings make a significant contribution by introducing perceived authenticity as a novel construct of interest for research on real-world connections in mathematics education. However, the study also shows that the basis of this significant perception of authenticity is largely unclear. Qualitative research into the sources of this perception is required so that one can better understand and address the construct theoretically.

(3) While Vos (2018) criticizes the indiscriminate use of the term "authentic tasks", this critique can be further extended. It is not only problematic to label tasks as "authentic" in a general sense, but also to assume that students uniformly perceive tasks as authentic. Rather, the heterogeneity of authenticity perception should be taken seriously and understood as a subjective construct that varies.

(4) In the practical development of reality-based large-scale test items, authenticity is repeatedly cited as a key criterion (OECD, 2023; Tout et al., 2017). The study results contribute to this discussion by highlighting that authenticity should not only be considered as a content-validity criterion, ensuring that aspects of reality-based test items correspond to the real world. In addition, consideration should be given to how students perceive this authenticity, as perceived authenticity has been shown to be relevant for emotional-motivational experiences during test item processing. For item development, this suggests that perceived authenticity could be considered during item review or piloting, for instance by examining whether students recognize the intended factual authenticity of item aspects. Further research on how perceived authenticity emerges may provide a basis for deriving concrete design principles aimed at supporting students' perception of authenticity.

Author contributions: DS: conceptualization, formal analysis, investigation, methodology, writing – original draft; **MB:** conceptualization, methodology, writing – review & editing; **LA:** conceptualization, investigation, writing – review & editing; **KR:** conceptualization, methodology, writing – review & editing. All authors agreed with the results and conclusions.

Funding: This publication was funded by the Open Access Publication Fund of Leuphana University Lüneburg.

Ethical statement: The authors declared that the study was conducted in accordance with the guidelines for safeguarding good research practice of the German Research Foundation. The study was approved by the relevant school supervisory authorities, the Regionales Landesamt für Schule und Bildung Lüneburg and the Ministerium für Allgemeine und Berufliche Bildung, Wissenschaft, Forschung und Kultur in Schleswig-Holstein. Participation was voluntary, and students could decline or withdraw without disadvantage. Students and their legal guardians were informed about the study and written informed consent was obtained. All data were analyzed in pseudonymized form.

AI statement: DeepL and ChatGPT were used solely for linguistic editing during the preparation of this manuscript. No AI tools were used to generate its scientific content.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

Amland, T., Grande, G., Scherer, R., Lervåg, A., & Melby-Lervåg, M. (2025). Cognitive factors underlying mathematical skills: A systematic review and meta-analysis. *Psychological Bulletin, 151*(1), 88-129. https://doi.org/10.1037/bul0000457

Bahar, A., & Maker, C. J. (2015). Cognitive backgrounds of problem solving: A comparison of open-ended vs. closed mathematics problems. *Eurasia Journal of Mathematics, Science and Technology Education, 11*(6), 1531-1546. https://doi.org/10.12973/eurasia.2015.1410a

Betz, A. (2018). Der Einfluss der Lernumgebung auf die (wahrgenommene) Authentizität der linguistischen Wissenschaftsvermittlung und das situationale Interesse von Lernenden [The influence of the learning environment on learners' (perceived) authenticity of science communication and on their situational interest]. *Unterrichtswissenschaft, 46*, 261-278. https://doi.org/10.1007/s42010-018-0021-0

Betz, A., Flake, S., Mierwald, M., & Vanderbeke, M. (2016). Modelling authenticity in teaching and learning contexts: A contribution to theory development and empirical investigation of the construct. In C.-K. Looi, J. Polman, U. Cress, & P. Reimann (Eds.), *Transforming learning, empowering learners: The International Conference of the Learning Sciences (ICLS) 2016* (pp. 815-818). International Society of the Learning Sciences.

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment, 31*(12), 1412-1427. https://doi.org/10.1037/pas0000626

Dowley, M., & Rice, S. (2022). Comparing student motivations for and emotional responses to national standardised tests and internal school tests: The devil in the detail. *Australian Journal of Education, 66*(1), 92-110. https://doi.org/10.1177/00049441211061889

Drüke-Noe, C. (2012). Können Lernstandserhebungen einen Beitrag zur Unterrichtsentwicklung leisten [Can assessments of learning progress contribute to the development of teaching methods]? In W. Blum, R. Borromeo Ferri, & K. Maaß (Eds.), *Mathematikunterricht im Kontext von Realität, Kultur und Lehrerprofessionalität* (pp. 284-293). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-8348-2389-2_29

Eckhardt, T. (2021). *The education system in the Federal Republic of Germany 2019/2020: A description of the responsibilities, structures and developments in education policy for the exchange of information in Europe*. KMK.

Enders, C. K. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological Methods, 8*(3), 322-337. https://doi.org/10.1037/1082-989X.8.3.322

Finger, L., van den Bogaert, V., Fleischer, J., Raimann, J., Sommer, K., & Wirth, J. (2022). Das Schülerlabor als Ort authentischer Wissenschaftsvermittlung? Entwicklung und Validierung eines Fragebogens zur Erfassung der Authentizitätswahrnehmung der Wissenschaftsvermittlung im Schülerlabor [Out-of-school labs as places for authentic science education? Construction and validation of a questionnaire for measuring the perception of authenticity of science education in out-of-school labs]. *Zeitschrift für Didaktik der Naturwissenschaften, 28*, Article 2. https://doi.org/10.1007/s40573-022-00139-4

Frenzel, A. C., Pekrun, R., Dicke, A.-L., & Goetz, T. (2012). Beyond quantitative decline: Conceptual shifts in adolescents' development of interest in mathematics. *Developmental Psychology, 48*(4), 1069-1082. https://doi.org/10.1037/a0026895

Geiger, V., Goos, M., & Forgasz, H. (2015). A rich interpretation of numeracy for the 21st century: A survey of the state of the field. *ZDM Mathematics Education, 47*, 531-548. https://doi.org/10.1007/s11858-015-0708-1

Grotjahn, R. (Ed.). (2010). *The C-test: Contributions from current research*. Peter Lang.

Gulikers, J. T. M., Bastiaens, T. J., & Martens, R. L. (2005). The surplus value of an authentic learning environment. *Computers in Human Behavior, 21*(3), 509-521. https://doi.org/10.1016/j.chb.2004.10.028

Hartmann, L. M., & Schukajlow, S. (2021). Interest and emotions while solving real-world problems inside and outside the classroom. In F. K. S. Leung, G. A. Stillman, G. Kaiser, & K. L. Wong (Eds.), *Mathematical modelling education in East and West. International perspectives on the teaching and learning of mathematical modelling* (pp. 177-188). Springer. https://doi.org/10.1007/978-3-030-66996-6_13

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision (KFT 4-12+R)* [Cognitive abilities test for grades 4 to 12, revision (KFT 4-12+R)]. Beltz Testgesellschaft.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*(2), 111-127. https://doi.org/10.1207/s15326985ep4102_4

Hohrath, S., Aßmann, S., Krabbe, H., & Opfermann, M. (2024). Students' perceived authenticity and understanding of authentic research while experimenting in a non-formal learning setting. *European Journal of Psychology of Education, 39*, 3325-3349. https://doi.org/10.1007/s10212-024-00810-z

International Association for the Evaluation of Educational Achievement. (2020). TIMSS – Trends in international mathematics and science study: Monitoring mathematics and science achievement in a global context. *IEA*. https://www.iea.nl/sites/default/files/2020-03/TIMSSGeneralflyer_0.pdf

Kanefke, J., & Schukajlow, S. (2024). I find this task interesting, so do you? Preservice teachers' judgments of students' enjoyment, boredom, and situational interest regarding tasks with and without a connection to reality. *Journal of Mathematics Teacher Education, 27*, 499-520. https://doi.org/10.1007/s10857-023-09581-8

Knabbe, A., Leiss, D., & Ehmke, T. (2025). Reality-based tasks with complex-situations: Identifying sociodemographic and cognitive factors for solution. *International Journal of Science and Mathematics Education, 23*, 97-120. https://doi.org/10.1007/s10763-024-10463-5

Knogler, M., Harackiewicz, J. M., Gegenfurtner, A., & Lewalter, D. (2015). How situational is situational interest? Investigating the longitudinal structure of situational interest. *Contemporary Educational Psychology, 42*, 39-50. https://doi.org/10.1016/j.cedpsych.2015.08.004

Krawitz, J., Schukajlow, S., Yang, X., & Geiger, V. (2025). A systematic review of international perspectives on mathematical modelling: Modelling goals and task characteristics. *ZDM Mathematics Education, 57*, 193-212. https://doi.org/10.1007/s11858-025-01683-2

Krug, A., & Schukajlow, S. (2013). Problems with and without connection to reality and students' task-specific interest. In A. M. Lindmeier & A. Heinze (Eds.), *Proceedings of the 37th Conference of the International Group for the Psychology of Mathematics Education* (vol. 3, pp. 209-216). PME.

Leiss, D., Plath, J., & Schwippert, K. (2019). Language and mathematics – Key factors influencing the comprehension process in reality-based tasks. *Mathematical Thinking and Learning, 21*(2), 131-153. https://doi.org/10.1080/10986065.2019.1570835

Lowrie, T. (2011). "If this was real": Tensions between using genuine artefacts and collaborative learning in mathematics tasks. *Research in Mathematics Education, 13*(1), 1-16. https://doi.org/10.1080/14794802.2011.550707

Mahler, N., Kölm, J., & Werner, B. (2020). Entwicklung von Mathematiktestaufgaben für Schülerinnen mit einem sonderpädagogischen Förderbedarf im Lernen – Konzeption und empirische Ergebnisse [Development of mathematics test items for students with special educational needs in learning – Conception and empirical results]. In C. Gresch, P. Kuhl, M. Grosche, C. Sälzer, & P. Stanat (Eds.), *Schülerinnen mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen* (pp. 109-146). Springer. https://doi.org/10.1007/978-3-658-27608-9_5

Mang, J., Seidl, L., Schiepe-Tiska, A., Tupac-Yupanqui, A., Ziernwald, L., Doroganova, A., Weis, M., Diedrich, J., Heine, J.-H., González Rodríguez, E., & Reiss, K. (2021). *PISA 2018. Skalenhandbuch: Dokumentation der Erhebungsinstrumente* [PISA 2018. Scales handbook: Documentation of the survey instruments]. Waxmann. https://doi.org/10.31244/9783830994961

Nachtigall, V., & Rummel, N. (2021). Investigating students' perceived authenticity of learning activities in an out-of-school lab for social sciences: A replication study. *Instructional Science, 49*, 779-810. https://doi.org/10.1007/s11251-021-09556-3

Nachtigall, V., Keuschnig, A., Behrendt, L., & Brune, L. (2018a). Authentic learning and teaching in an out-of-school lab—First steps towards empirical investigation of a theoretical model. In J. Kay, & R. Luckin (Eds.), *Rethinking learning in the digital age: Making the learning sciences count, 13th International Conference of the Learning Sciences (ICLS) 2018* (vol. 2, pp. 1061-1064). International Society of the Learning Sciences.

Nachtigall, V., Rummel, N., & Serova, K. (2018b). Authentisch ist nicht gleich authentisch – Wie Schülerinnen und Schüler die Authentizität von Lernaktivitäten im Schülerlabor einschätzen [Authentic does not equal authentic – How students evaluate the authenticity of learning activities in an out-of-school lab]. *Unterrichtswissenschaft, 46*, 299-319. https://doi.org/10.1007/s42010-018-0020-1

Nachtigall, V., Shaffer, D., & Rummel, N. (2022). Stirring a secret sauce: A literature review on the conditions and effects of authentic learning. *Educational Psychology Review, 34*, 1479-1516. https://doi.org/10.1007/s10648-022-09676-3

National Assessment Governing Board. (2021). Mathematics framework for the 2026 national assessment of educational progress. *NAGB*. https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/mathematics/2026-math-frameowork/NAEP-2026-Mathematics-Framework-Combined.pdf

Niss, M., & Blum, W. (2020). *The learning and teaching of mathematical modelling*. Routledge. https://doi.org/10.4324/9781315189314

OECD. (2021). *The assessment frameworks for cycle 2 of the Programme for the International Assessment of Adult Competencies*. OECD Publishing. https://doi.org/10.1787/4bc2342d-en

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

Palm, T. (2008). Impact of authenticity on sense making in word problem solving. *Educational Studies in Mathematics, 67*, 37-58. https://doi.org/10.1007/s10649-007-9083-3

Palm, T. (2009). Theory of authentic task situations. In L. Verschaffel, B. Greer, W. Van Dooren, & S. Mukhopadhyay (Eds.), *Words and worlds: Modelling verbal descriptions of situations* (pp. 3-19). Sense Publishers. https://doi.org/10.1163/9789087909383_002

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*, 315-341. https://doi.org/10.1007/s10648-006-9029-9

Prediger, S., Wilhelm, N., Büchter, A., Gürsoy, E., & Benholz, C. (2018). Language proficiency and mathematics achievement. *Journal für Mathematik-Didaktik, 39*, 1-26. https://doi.org/10.1007/s13138-018-0126-3

Quarder, J., Schukajlow, S., Krawitz, J., Wiehe, K., & Rakoczy, K. (2025). Experiences of autonomy and competence and their reciprocal relationship: The roles of prior performance and open modelling problems in mathematics teaching. *International Journal of Science and Mathematics Education, 23*, 4127-4151. https://doi.org/10.1007/s10763-025-10618-y

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J., & Schiefele, U. (2012). *Fragebogenskalen (Skalenkollektion): PISA-I* [Questionnaire scales (scale collection): PISA-I]. Forschungsdatenzentrum Bildung am DIPF. https://doi.org/10.7477/15:287:1

Schlüter, D., & Besser, M. (2024). Assessing authenticity in modelling test items: Deriving a theoretical model. *Frontiers in Education, 9*, Article 1343510. https://doi.org/10.3389/feduc.2024.1343510

Schlüter, D., Besser, M., Greefrath, G., & Vos, P. (2026). Empirical analysis of authenticity in reality-based mathematical test items. *Research in Mathematics Education*, 1-22. https://doi.org/10.1080/14794802.2026.2692608

Schukajlow, S., Krawitz, J., Kanefke, J., & Rakoczy, K. (2022). Interest and performance in solving open modelling problems and closed real-world problems. In C. Fernández, S. Llinares, A. Gutiérrez, & N. Planas (Eds.), *Proceedings of the 45th Conference of the International Group for the Psychology of Mathematics Education* (vol. 3, pp. 403-410). PME.

Schukajlow, S., Leiss, D., Pekrun, R., Blum, W., Müller, M., & Messner, R. (2012). Teaching methods for modelling problems and students' task-specific enjoyment, value, interest and self-efficacy expectations. *Educational Studies in Mathematics, 79*, 215-237. https://doi.org/10.1007/s10649-011-9341-2

Schukajlow, S., Rakoczy, K., & Pekrun, R. (2023). Emotions and motivation in mathematics education: Where we are today and where we need to go. *ZDM Mathematics Education, 55*, 249-267. https://doi.org/10.1007/s11858-022-01463-2

Schüttler, T., Watzka, B., Girwidz, R., & Ertl, B. (2021). Die Wirkung der Authentizität von Lernort und Laborgeräten auf das situationale Interesse und die Relevanzwahrnehmung beim Besuch eines naturwissenschaftlichen Schülerlabors [Effects of an authentic location and laboratory equipment for the situational interest and the perception of content relevance when visiting an out-of-school science lab]. *Zeitschrift für Didaktik der Naturwissenschaften, 27*, 109-125. https://doi.org/10.1007/s40573-021-00128-z

Siller, H. S., & Greefrath, G. (2020). Modelling tasks in central examinations based on the example of Austria. In G. A. Stillman, G. Kaiser, & C. E. Lampen (Eds.), *Mathematical modelling education and sense-making. International perspectives on the teaching and learning of mathematical modelling* (pp. 383-392). Springer. https://doi.org/10.1007/978-3-030-37673-4_33

Siller, H. S., Geiger, V., & Kaiser, G. (2024). Researching mathematical modelling education in disruptive times – An introduction. In H. S. Siller, V. Geiger, & G. Kaiser (Eds.), *Researching mathematical modelling education in disruptive times. International perspectives on the teaching and learning of mathematical modelling* (pp. 3-11). Springer. https://doi.org/10.1007/978-3-031-53322-8_1

Stacey, K. (2015). The real world and the mathematical world. In K. Stacey, & R. Turner (Eds.), *Assessing mathematical literacy* (pp. 57-84). Springer. https://doi.org/10.1007/978-3-319-10121-7_3

Stamer, I., David, M. A., Höffler, T., Schwarzer, S., & Parchmann, I. (2021). Authentic insights into science: Scientific videos used in out-of-school learning environments. *International Journal of Science Education, 43*(6), 868-887. https://doi.org/10.1080/09500693.2021.1891321

Sundtjønn, T. (2025). Complexities when students authenticate aspects of a task in vocational mathematics education. In M. Bosch, G. Bolondi, S. Carreira, C. Spagnolo, & M. Gaidoschik (Eds.), *Proceedings of the 14th Congress of the European Society for Research in Mathematics Education* (pp. 1238-1245). ERME.

Tanaka, A., & Murayama, K. (2014). Within-person analyses of situational interest and boredom: Interactions between task-specific perceptions and achievement goals. *Journal of Educational Psychology, 106*(4), 1122-1134. https://doi.org/10.1037/a0036659

Tout, D., & Spithill, J. (2015). The challenges and complexities of writing items to test mathematical literacy. In K. Stacey, & R. Turner (Eds.), *Assessing mathematical literacy* (pp. 145-171). Springer. https://doi.org/10.1007/978-3-319-10121-7_7

Tout, D., Coben, D., Geiger, V., Ginsburg, L., Hoogland, K., Maguire, T., Thomson, S., & Turner, R. (2017). *Review of the PIAAC numeracy assessment framework: Final report*. Australian Council for Educational Research. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1033&context=transitions_misc

Tran, D., Nguyen, D. T., Nguyen, A. T. T., Nguyen, G. N. T., & Ta, P. M. (2019). Bridging to mathematical modelling: Vietnamese students' response to different levels of authenticity in contextualized tasks. *International Journal of Mathematical Education in Science and Technology, 51*(6), 893-912. https://doi.org/10.1080/0020739X.2019.1648890

Verschaffel, L., Greer, B., & De Corte, E. (2000). *Making sense of word problems*. Swets and Zeitlinger.

Vos, P. (2011). What is 'authentic' in the teaching and learning of mathematical modelling? In G. Kaiser, W. Blum, R. Borromeo Ferri, & G. Stillman (Eds.), *Trends in teaching and learning of mathematical modelling. International perspectives on the teaching and learning of mathematical modelling* (pp. 713-722). Springer. https://doi.org/10.1007/978-94-007-0910-2_68

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

Vos, P., Devesse, T. G., & Pinto, A. A. R. (2007). Designing mathematics lessons in Mozambique: Starting from authentic resources. *African Journal of Research in Mathematics, Science and Technology Education, 11*(2), 51-66. https://doi.org/10.1080/10288457.2007.10740621

Weiss, L., & Müller, A. (2015). The notion of authenticity in the PISA units in physical science: An empirical analysis. *Zeitschrift für Didaktik der Naturwissenschaften, 21*, 87-97. https://doi.org/10.1007/s40573-015-0025-9

Zachrich, L., Wagner, W., Bertram, C., & Trautwein, U. (2024). Really? It depends! How authentic learning material affects involvement with personal stories of the past. *Learning and Instruction, 92*, Article 101921. https://doi.org/10.1016/j.learninstruc.2024.101921

Živković, M., Pellizzoni, S., Doz, E., Cuder, A., Mammarella, I., & Passolunghi, M. C. (2023). Math self-efficacy or anxiety? The role of emotional and motivational contribution in math performance. *Social Psychology of Education, 26*, 579-601. https://doi.org/10.1007/s11218-023-09760-8

APPENDIX A: QUESTIONNAIRE ITEMS USED IN THE STUDY

Table A1. Socio-demographic characteristics

Variable	*German version*	*English translation*
Age	Bitte gib dein Alter an.	Please state your age.
Gender	Bitte gib dein Geschlecht an.	Please state your gender.
Migration background (taken from Mang et al., 2021)	1) In welchem Land wurdest du geboren? 2) In welchem Land wurde deine Mutter geboren? 3) In welchem Land wurde dein Vater geboren?	In which country were you born? In which country was your mother born? In which country was your father born?
HEDRES (home educational resources, taken from Mang et al., 2021)	Welche der folgenden Dinge gibt es bei dir zuhause? 1) Einen Schreibtisch zum Lernen 2) Einen ruhigen Platz zum Lernen 3) Einen Computer, den du für Schularbeiten verwenden kannst 4) Lernsoftware 5) Bücher, die dir bei Schularbeiten helfen 6) Technische Nachschlagewerke 7) Ein Wörterbuch	Which of the following things do you have at home? 1) A desk for studying 2) A quiet place to study 3) A computer that you can use for schoolwork 4) Educational software 5) Books to help with schoolwork 6) Technical reference works 7) A dictionary

Table A2. General emotional-motivational variables related to mathematics

Variable	*German version*	*English translation*
Mathematics-related interest (taken from Frenzel et al., 2012)	1) Ich interessiere mich für Mathematik. 2) Ich lese gerne Bücher und löse Knobelaufgaben, die mit Mathematik zu tun haben. 3) Mathematik gehört zu meinen Lieblingsbeschäftigungen. 4) Ich finde die Dinge, mit denen wir uns in Mathematik beschäftigen, oft spannend. 5) Nach einer Mathestunde bin ich oft neugierig, was wir in der nächsten Stunde machen werden. 6) Ich würde gerne mehr über einige der Dinge erfahren, die wir in unserem Mathematikunterricht behandeln.	1) I am interested in mathematics. 2) I like to read books and solve brain teasers related to mathematics. 3) Doing mathematics is one of my favorite activities. 4) I often find that the things we deal with in mathematics are really exciting. 5) After a math class, I am often curious about what we are going to do in the next lesson. 6) I would like to find out much more about some of the things we deal with in our mathematics class.
Mathematics-related self-efficacy (taken from Ramm et al., 2012)	1) In Mathematik bin ich sicher, dass ich auch den schwierigsten Stoff verstehen kann. 2) Ich bin überzeugt, dass ich auch den kompliziertesten Stoff, den der Lehrer in Mathematik vorstellt, verstehen kann. 3) Ich bin überzeugt, dass ich in Hausaufgaben und Prüfungen in Mathematik gute Leistungen erzielen kann. 4) Ich bin überzeugt, dass ich die Fertigkeiten, die in Mathematik unterrichtet werden, beherrschen kann.	1) When it comes to mathematics, I'm sure I can understand even the most difficult topics. 2) I am convinced that I can understand even the most complicated topics the mathematics teacher covers. 3) I am confident that I can achieve good results in mathematics homework and exams. 4) I am convinced that I can master the skills taught in mathematics.
Mathematics-related self-concept (taken from Ramm et al., 2012)	1) Ich bin einfach nicht gut in Mathematik *(umgepolt)*. 2) Im Fach Mathematik bekomme ich gute Noten. 3) In Mathematik lerne ich schnell. 4) Ich war schon immer überzeugt, dass Mathematik eines meiner besten Fächer ist. 5) Im Mathematikunterricht verstehe ich sogar die schwierigsten Aufgaben.	1) I'm just not very good at mathematics *(reverse-coded)*. 2) I get good marks in mathematics. 3) I learn mathematics quickly. 4) I've always been convinced that mathematics is one of my strongest subjects. 5) In mathematics lessons, I can even understand the most difficult problems.
Mathematics anxiety (taken from Ramm et al., 2012)	1) Ich mache mir oft Sorgen, dass es für mich im Mathematikunterricht schwierig wird. 2) Ich bin sehr angespannt, wenn ich Mathematikhausaufgaben machen muss. 3) Beim Lösen von Aufgaben in Mathematik werde ich sehr nervös. 4) Ich fühle mich beim Lösen von Mathematikaufgaben hilflos. 5) Ich mache mir Sorgen, dass ich in Mathematik schlechte Noten bekomme.	1) I often worry that I'll find mathematics lessons difficult. 2) I feel very tense when I have to do my mathematics homework. 3) I get very nervous when I'm doing mathematics problems. 4) I feel at a loss when it comes to solving mathematics problems. 5) I'm worried that I'll get poor marks in mathematics.

Table A3. Test item-specific variables

Variable	*German version*	*English translation*
Perceived authenticity (newly developed)	1) In der Einleitung steht, dass [Beschreibung]. Dieses Ereignis/diese Ereignisse stammen aus dem echten Leben. 2) Die Frage aus der Aufgabe stellt man sich auch im echten Leben. 3) Bei der Bearbeitung der Aufgabe geht man mathematisch so vor, wie man es im echten Leben auch macht.	1) The introduction states that [description of the event(s)]. This event/these events is/are based on real life. 2) The question posed in the task is one that arises in real life as well. 3) The mathematical approach to solving this problem is the same as in real life.
Situational interest (based on Kanefke & Schukajlow, 2024)	Ich fand es interessant, die Aufgabe zu bearbeiten.	I found working on the task interesting.
Boredom (based on Kanefke & Schukajlow, 2024)	Ich fand es langweilig, die Aufgabe zu bearbeiten.	I was bored working on the task.
Enjoyment (based on Kanefke & Schukajlow, 2024)	Ich hatte Spaß dabei, die Aufgabe zu bearbeiten.	I enjoyed working on the task.